



저작자표시 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#) 

Auto-Segmentation of Target Volume and Organs-at-risks for Radiotherapy in Breast Cancer patients

Jee Suk Chang

Department of Medicine

The Graduate School, Yonsei University

Auto-Segmentation of Target Volume and Organs-at-risks for Radiotherapy in Breast Cancer patients

Directed by Professor Ki Chang Keum

Doctoral Dissertation
submitted to the Department of Medicine,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree
of Doctor of Philosophy in Medical Science

Jee Suk Chang

June 2021

This certifies that the Doctoral
Dissertation of Jee Suk Chang is
approved.

Thesis Supervisor: Ki Chang Keum

Thesis Committee Member#1 : Jin Sung Kim

Thesis Committee Member#2 : Seho Park

Thesis Committee Member#: Hwiyoung Kim

Thesis Committee Member#4: Juree Kim

The Graduate School
Yonsei University

June 2021

ACKNOWLEDGEMENTS

I would like to extend immeasurable appreciation and deepest gratitude for many precious people's help and support. First, I would like to express my appreciation and great respect to my supervisor, Prof. Ki Chang Keum, for guiding me in every step of this thesis, research, and clinical practice. He has been a tremendous source of professional inspiration to me, as a great mentor with enthusiasm and wisdom.

I would also like to show gratitude to the members of advisory committee: Jin Sung Kim, Seho Park, Hwiyoung Kim, and Juree Kim for their critical comments and valuable suggestions.

I express sincere thanks to Prof. Yong Bae Kim for his guidance in all research in breast cancer areas. I also appreciate Seung Yeun Chung, Min Seo Choi, Hwa Kyung Byun, Nalee Kim, Yongjin Chang, Jaehee Chung and Jin Sung Kim for their invaluable assistance in artificial intelligence studies.

I also thank Hong In Yoon, Yeona Cho, and Kyung Hwan Kim, the young members of our department. They always give me motivation and inspiration. Lastly, I send my warmest regards to my family, Jihyun Park, Teo Chang, and Seah Chang.

For the rest of my life, I will dedicate my best effort to cancer treatment and research for the cancer patients.

<TABLE OF CONTENTS>

ABSTRACT	1
I. INTRODUCTION	3
II. MATERIALS AND METHODS	7
1. COMPARISON with ABAS	7
2. INVESTIGATION OF CLINICAL USEFULNESS	11
3. EXTERNAL VALIDATION	16
III. RESULTS	19
1. COMPARISON with ABAS	19
2. INVESTIGATION OF CLINICAL USEFULNESS	28
3. EXTERNAL VALIDATION	36
IV. DISCUSSION	51
V. CONCLUSION	61
REFERENCES	62
ABSTRACT (IN KOREAN)	68
PUBLICATION LIST	70

LIST OF FIGURES

Fig. 1. The schematic of the proposed FCDN.	10
Fig. 2. Schematic of the proposed convolutional neural network architecture (U-Net with EfficientNet-B0)	13
Fig. 3. Examples of a) CTV, b) OAR, and c) heart segmentation results of DLBAS based on FCDN and ABAS by MIM and Mirada compared against ground-truth manual contours	19
Fig. 4. Box-plots of Dice Similarity Coefficients (DSC) and 95% Hausdorff Distance (HD) in the a) CTVs, b) OARs, and c) Heart structures obtained from Mirada, MIM, and DLBAS based on FCDN using the manual contours as reference. ...	25
Fig. 5. Difference in dice similarity coefficients (Δ DSC) between contrast and non-contrast test sets obtained from DLBAS based on FCDN and ABAS	26
Fig. 6. Difference in 95% Hausdorff distance (Δ HD) between contrast and non-contrast test sets obtained from fully convolutional DenseNet (FCDN) and atlas-based auto-segmentation (ABAS).	27
Fig. 7. Example of deep learning-based auto-segmentation (green) and manual contours (red).	28
Fig. 8. Comparison of dose-volume histograms with average dosimetric values of manual contours (solid line) and auto-segmentation contours (dotted line) for patients who received whole breast RT only (A) or that with regional node	

irradiation (B).	30
Fig. 9. Examples of deep learning-based auto-contour without correction.....	36
Fig. 10. Dice similarity coefficient (A) and Haudorff distance values (B) according to the organ-at-risks, comparing manual contours, corrected-auto-contours, and auto-contours. For sensitivity analyses, contouring metrics were obtained by comparing each contour with the second-best contour.	37
Fig. 11. Radar graphs showing mean DSC value of each participant according to the organs: (A) manual contours, (B) corrected-auto-contours. DSC values of corrected-auto-contours were more homogeneous than those of manual contours meaning that inter-physician variability was reduced.	39
Fig. 12. Radar graphs showing mean DSC value of each participant according to the organs: (A) manual contours, (B) corrected-auto-contours. DSC values of corrected-auto-contours were more homogeneous than those of manual contours meaning that inter-physician variability was reduced. For sensitivity analyses, contouring metrics were obtained by comparing each contour with the second-best contour.....	39
Fig. 13. Examples of manual and corrected-auto-contours of all experts: (A) breast contours showing that inter-physician variability is mostly seen in lateral and anterior borders of	

breasts and is reduced with an aid of auto-contouring	40
Fig. 14. Contouring time comparing manual contouring and correcting auto-contours: (A) total contouring time of all 9 organ-at-risks of each expert, (B) contouring time of each organ-at-risks.	41

LIST OF TABLES

Table 1. Inter-observer variability of manual contours of organs-at-risk and target volumes	16
Table 2. Comparison of average DSC, HD and their significance for CTV segmentation of the patients in the test set	21
Table 3. Comparison of average DSC, HD and their significance for OAR segmentation of the patients in the test set	22
Table 4. Comparison of average DSC, HD, and their significance for heart segmentation of the patients in the test set	24
Table 5. Comparison of deep learning auto-segmentation and manual contours of organs-at-risk and target volumes	32
Table 6. Dosimetric outcomes for manual and auto-segmented contours.	34
Table 7. Summary of Dice Similarity Coefficient and Hausdorff distance	42
Table 8. Summary of Dice Similarity Coefficient and Hausdorff distance	43

Table 9. Summary of Dice Similarity Coefficient and Hausdorff distance for sensitivity analyses.	45
Table 10. The DSC and HD values of all OARs of experts' manual contours and an auto-contour, listed from the best to the lowest performance.	47
Table 11. Total contouring time for 9 organ-at-risks of each patient	48
Table 12. Time for manual contouring according to the organ-at-risks.....	49
Table 13. Time for correcting auto-contours according to the organ-at-risks.....	41

ABSTRACT

Auto-Segmentation of Target Volume and Organs-at-risks for Radiotherapy in Breast Cancer patients

Jee Suk Chang

*Department of Medicine
The Graduate School, Yonsei University*

(Directed by Professor Ki Chang Keum)

Background and purpose: In breast cancer patients receiving radiotherapy, accurate target delineation and reduction of radiation doses to the nearby normal organs is important. However, manual clinical target volume (CTV) and organs-at-risk (OAR) segmentation for treatment planning increases physicians' workload and inter-physician variability considerably. In this study, we first evaluated the feasibility of a deep learning-based auto-segmentation (DLBAS) in comparison to atlas-based segmentation solutions (ABAS) for breast radiation therapy. Secondly, we evaluated the clinical utility of proposed-DLBAS from a clinician's perspective. Lastly, external validation was conducted.

Methods and materials: CTVs and OARs were generated by one expert on planning CT scans of breast cancer patients. Auto-contours were generated using convolutional neural network algorithm. First, accuracy of DLBAS was compared with ABAS using Dice similarity coefficient (DSC) and 95% Hausdorff distance (HD). Secondly, additional qualitative scoring of DLBAS and dose-volume histograms with dosimetric parameters were analyzed. Lastly, 11 experts from two institutions were asked to participate in this external validation. Each contour of DLBAS and 11 manual contours were compared with the best manual contour, which was selected by independent committee.

Results: Compared to ABAS, the proposed DLBAS model yielded more consistent results and the highest average Dice similarity constant and lowest Hausdorff distances, especially CTVs and the substructures of the heart. ABAS showed limited performance in soft-tissue-based regions, such as the esophagus, cardiac arteries, and smaller CTVs. The results of sensitivity analysis between contrast and non-contrast CT test sets showed little difference in the performance of DLBAS and conversely, a large discrepancy for ABAS. Secondly, qualitative subjective scoring showed that the results were acceptable for all CTVs and OARs, with a median score of at least 8 (possible range: 0–10) for (1) the differences between manual and auto-segmented contours and (2) the extent to which auto-segmentation would assist physicians in clinical practice. The differences in dosimetric parameters between the auto-segmented and manual contours were minimal. In external validation, Total mean time for 9 OARs was 37 ± 20 min for manual and 6 ± 5 min for corrected-auto-contours. Among the DSC of experts' manual contours and an auto-contour, DSC of an auto-contour ranked the second place and HD ranked the first place. Among manual OARs, breast contours had the largest variations, which were most significantly improved with an aid of ACS.

Conclusions: The feasibility of deep learning-based auto-segmentation in breast RT planning was demonstrated. Although deep learning-based auto-segmentation cannot be a substitute for radiation oncologists, it is a useful tool with excellent potential in assisting radiation oncologists in the future.

Key words: breast cancer, auto-segmentation, radiation therapy, artificial intelligence.

Auto-Segmentation of Target Volume and Organs-at-risks for Radiotherapy in Breast Cancer patients

Jee Suk Chang

*Department of Medicine
The Graduate School, Yonsei University*

(Directed by Professor Ki Chang Keum)

I. INTRODUCTION

Breast cancer is the most common cancer among women in many countries, accounting for 1 in 4 newly diagnosed cancers.¹ Moreover, it is the leading cause of cancer deaths in women in over 100 countries. As an integral part of the curative treatment for patients with breast cancer, the clinical utilization of radiation therapy (RT) has increased in recent decades. Thus, in this era of three-dimensional conformal and intensity-modulated RT, precise delineation of the clinical target volume (CTVs) and the organ at risk (OAR) has become vital. The requirements for accurate delineation of delicate target volumes for regional node irradiation have increased with the availability of recent data from the EORTC 22922-10925,² MA.20,³ and DBCG-IMN trials.⁴ We have previously reported that an individualised target volume is recommended even for patients in the modern management era.^{5,6} Because the majority of patients with breast cancer often survive for decades after receiving RT, they are at risk

of experiencing long-term adverse events, including lymphedema,⁷ radiation pneumonitis,⁸ hypothyroidism,⁹ and cardiotoxicity,¹⁰ which can substantially decrease quality of life.

Modern RT planning is a complex process that relies on computed tomography (CT)-based three-dimensional imaging as well as an expert team.¹¹ Based on CT simulations, radiation oncologists contour the relevant target volumes and surrounding normal structures and communicate with the dosimetrist the anticipated dosimetric goals that will deliver a therapeutic radiation dose to the target while sparing the OARs. In contrast to other primary malignancies such as lung and head & neck cancer, modern RT planning has not been commonly applied to breast cancer, in which conventional formulaic field-based planning and two-dimensional techniques were predominantly used¹². Currently, in RT planning, OARs and CTVs are manually segmented by radiation oncologists following international contour guidelines, such as the ESTRO and RTOG guidelines.¹³ However, given the typical number and complexity of the structures involved, delineation is a laborious and often time-consuming task. Additionally, as the outcome is highly dependent on the skill of the observer, a significant amount of inter-observer variation exists. A previous study showed that the contours from multiple observers had low structural overlap with volume variations of up to 60%, which would likely result in substantial variations in RT dosimetric planning.¹⁴ Quality issues and inter-physician variations of target volumes and OAR contours have been of particular concern arising from dummy runs, multi-institutional studies, individual case reviews and audit studies.^{15,16} Uncertainties regarding volume delineation and subsequent target and normal tissue doses may not only decrease the treatment efficacy, but also increase the complication risk.

Recognizing the limitations of the manual segmentation process, recent developments in auto-segmentation have gained significant attention for their potential application in routine clinical workflows. One solution that is currently available is atlas-based auto-segmentation (ABAS); to date, several commercial ABAS solutions have been released. There have been numerous studies evaluating ABAS for use in cancer sites such as the head and neck,¹⁷ prostate,¹⁸ and lungs.¹⁹ However, ABAS has several limitations, including insufficient contour results for structures with low contrast, slow image registration, and the need for additional correction time to improve segmentation accuracy.²⁰

More recently, with the increase in available computational power and the reduction of financial barriers, most research focus has shifted towards deep learning-based auto-segmentation (DLBAS) approaches.²¹ With recent advances in computing power, algorithms, and data collection, artificial intelligence (AI) is increasingly being used in health care to assist physicians. In radiation oncology, there are numerous areas in which AI is applicable, such as target and normal tissue segmentation, dose optimization, decision support systems, application of predictive models, and quality assurance ²²⁻²⁴. Auto-contouring tools have been adopted by an increasing number of physicians and have resulted in improved efficiency, particularly for OARs in head and neck cancer and target volume in prostate cancer ^{25,26}.

The first aim of this study is to evaluate commercially available ABAS and DLBAS methods used for delineating RT planning structures (CTVs and OARs) for breast cancer. As there is a paucity of data regarding the auto-segmentation of target volumes and OARs in breast RT planning, secondly, we attempted to train a DLBAS model for target volumes and OARs for breast cancer and evaluated its clinical utility from a clinician's perspective. Thirdly, we aimed to

externally validate the performance of the auto-contouring system, by comparing it with the manual contours of experts. We also examined whether the use of auto-contouring system in breast cancer radiotherapy would reduce the workload of radiation oncologists, promote accuracy of delineating OARs, and to reduce inter-physician variability.

II. MATERIALS AND METHODS

1. COMPARISON with ABAS

Sixty-two breast cancer patients, with a mix of left- and right-sided breast cancer, who received RT after breast-conserving surgery between May 1st, 2016 and May 1st, 2019 and underwent contrast-enhanced planning CT were selected for this study after Institutional Review Board approval. The CT scans were acquired on a Siemens Sensation Open (Siemens, Forchheim, Germany) and a Toshiba Aquilion (Toshiba Medical Systems, Japan), using the following common CT image acquisition protocols: 120 kVP, 3 mm CT slice thickness, and the patient in supine position with both arms up.

The expert contours were manually delineated by a single experienced radiation oncologist with over 10 years of experience who is currently treating over 500 breast cancer patients per year, all of whom are undergoing adjuvant RT with volumetric modulated arc therapy. For these patients, the planning contours included a full list of OARs and CTVs drawn following the ESTRO guidelines, to evaluate the performance of each software package in a realistic clinical workflow, which often involves the delineation of all of these structures. We included various types of CTVs, such as the axillary (AXL) Level 1-3, internal mammary (IMN), and supraclavicular lymph nodes (SCL). In the case of the supraclavicular nodes, we included both versions suggested by ESTRO (E) and RTOG (R) guidelines. In addition to other OARs (right and left lung, esophagus, spinal cord, and thyroid), we also evaluated the segmentation of heart substructures for further consideration of cardiotoxicity: right and left atrium, right and left ventricle, right coronary artery (RCA) and left anterior descending artery (LAD).

In addition, to test each algorithm's robustness to input data type, we conducted a sensitivity analysis by comparing the extent of change in Dice Similarity Coefficient (DSC) and 95% Hausdorff Distance (HD) of three segmentation models using 14 non-contrast CT scans. These scans were acquired using similar acquisition protocols as the contrast-enhanced CT sets and included all contour structures except for heart substructures.

Two commercial systems that perform user-defined ABAS— Mirada's Workflow Box (WFB, Mirada Medical, Ltd., Oxford, UK) and MIM Maestro (MIM Software Inc., Cleveland, OH)—were used to automatically segment target structures. We constructed an atlas library database for each software package that consists of the data from 35 patients and the corresponding expert contours.

In MIM, the first step in building an atlas was to assign a randomly selected reference or "template" subject. The remaining subjects were registered to the template one by one, along with the expert contours. Although MIM offers a tool to edit the registration alignment, to obtain a non-biased auto-segmentation and keep the experimental settings as consistent as possible, we did not intervene during registration and segmentation. The final step was the segmentation process itself. In MIM, under the "Atlas Segment" tool, we selected the contours and ran the segmentation with the following default settings: Number of Match = 1, Mirroring Enabled and Multicontour finalisation method = Majority Vote. Next, because a single atlas segmentation was selected, the algorithm automatically searched for the atlas subject that best matched the input CT. Then, expert contours of the atlas subject were deformed, registered, and transferred to the input CT, based on intensity and a freeform

cubic spline interpolation.

In Mirada, a workflow that linked the atlas created by the user and the segmentation operation was created that simply required selecting the input CT and assigning it to the workflow in a single click. As it functions like a black box, it is not possible to change settings in WFB. Also, unlike MIM, WFB does not require the assignment of template patients or any further user intervention. The construction of the library simply involved selecting CT scans and their corresponding structures. Once every subject was added, the atlas files were uploaded to the WFB server.

In this study, we developed a two-step of three-dimensional (3D) fully convolutional DenseNet (FCDN) to automatically contour the target structures in a semantic manner, as originally proposed by Jegou et al.²⁷ More detailed information regarding the two-step approach is included in the Supplementary data. The FCDN network was trained on an NVIDIA TITAN RTX GPU with Tensorflow in Python, using the same 35 patient scans as in Section 2.2, over 200 training epochs, with 13 patient scans used for validation.

As shown in Figure 1, the FCDN architecture is made up of dense blocks that resemble the residual blocks in a U-Net architecture. Following the convolution layer, the transition down layers consist of BN, RELU, 1×1 convolution, dropout ($p = 0.2$), and a 2×2 max pooling operation. The skip connection components represent the concatenation of the feature maps from the down-sampling path with those in the up-sampling path, thereby ensuring a high-resolution output. Finally, the transition up (TU) layers consist of 3×3 deconvolutions with a stride of 2 to progressively recover spatial resolution.

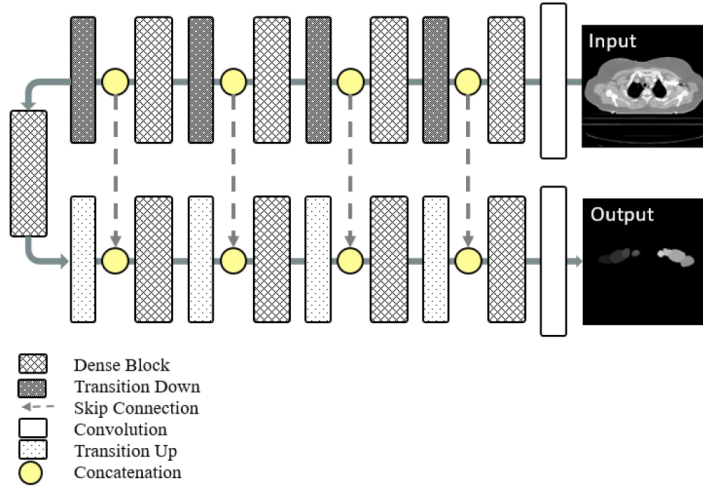


Fig. 1. The schematic of the proposed FCDN.

The accuracy of each segmentation method was assessed with 14 test patients using the DSC and HD. The manual contours delineated by a single expert (RO) were considered ground truth in this study, against which the ABAS and FCDN contours were compared. DSC is a metric that quantifies the closeness of the automated and expert contours, defined as double the overlap of the two contours divided by the sum of their individual volumes.

The range of the Dice scores is $[0, 1]$ where 1 indicates a perfect match between the two contours and 0 indicates no overlap at all. In this study, a Dice score of 0.75 was considered an acceptable match. HD describes the largest surface-to-surface separation among the 95th percentile of surface points of ground truth and segmentation. Similar procedures were repeated for the robustness analysis using 14 non-contrast CT test samples.

A pairwise t-test was conducted to determine if there was a statistically significant difference between the results from the different software packages.

Since there are three segmentation methods to compare, we adopted Bonferroni correction to address the multiple-comparison correction with $n = 3$ and the alpha value adjusted to 0.0167 (0.05/3). A p-value of less than 0.0167 was determined to be a rejection of the null hypothesis and therefore a statistically significant result.

2. INVESTIGATION OF CLINICAL USEFULNESS

It included 111 breast cancer patients who received adjuvant RT after breast-conserving surgery. Both left-sided and right-sided breast cancer patients were included. The median age of the patients was 51 years (range, 28–77 years) and the median body mass index was 22.5 kg/m² (range, 17.03–35.4 kg/m²). For T stage, 15 patients were Tis (14%), 60 patients were T1 (54%), 33 patients were T2 (30%), and 3 patients were T3 (3%). For N stage, 82 patients were N0 (74%), 26 patients were N1 (23%), and 3 patients were N2 (3%). RT field included whole breast only for 79 patients (71%) and WB with regional lymph nodes for 32 patients (29%). Both non-contrast ($n = 50$) and contrast-enhanced ($n = 61$) planning CT scans were used for manual delineation of CTVs and OARs. Planning CT scan (Somatom Sensation Open syngo CT 2009E, Siemens and Aquilion TSX-201A, Toshiba) was performed approximately two weeks prior to RT with a CT slice thickness of 3 mm. The setup position for all planning CT scans was the supine position with both arms held up using an arm support device (CIVICO). Contrast-enhanced planning CT was performed 1 min after administration of 80–90 mL intravenous contrast (iohexol, 84.11 g / 130 mL; depending on the patient's weight).

Previous contours used for patient treatment were not used in this study. For homogeneity, a single expert who is ESTRO teaching course certified and treats

approximately 550 breast cancer patients per year contoured the CTVs and OARs within 1 month, with the patients' clinical information blinded. The target volume consisted of CTVs of right and left breasts (CTVp_breast); axillary levels 1, 2, and 3 (CTVn_L1, L2, L3); internal mammary chain (CTVn_IMN); and lymph node level 4 (CTVn_L4), which is supraclavicular lymph node delineated according to the ESTRO guidelines²⁸. In our study, we included interpectoral nodes mentioned in the ESTRO guidelines in CTVn_L2. In addition, the supraclavicular lymph nodes were additionally delineated according to the RTOG guidelines (CTVn_SCL RTOG).²⁹ The OARs included the heart, right and left lungs, esophagus, spinal cord, and thyroid.³⁰

To segment the CTVs and OARs, a convolutional neural network (CNN) was used, which combined a U-Net with EfficientNet-B0 as the backbone (Figure 2).³¹ In EfficientNet, 3D convolutional layers are used to exploit the 3D structural information.³² For inputs of the CNN, all cases were resampled to a voxel spacing of 1.0×1.0×3.0 mm³ and then the image intensity values of a truncated range of [-160, 240] were normalized into the range of [0, 1]. Owing to GPU memory limitations, the CNN was trained in the patch level, with a size of 128×128×64. Furthermore, we trained the CNN with the sum of cross-entropy and dice loss, and we used RMSprop optimizer with an initial learning rate of 5×10⁻⁴ and weight decay of 10⁻⁴.³³ During training, we applied data augmentation techniques such as scaling, flip, and rotation to all training patches.

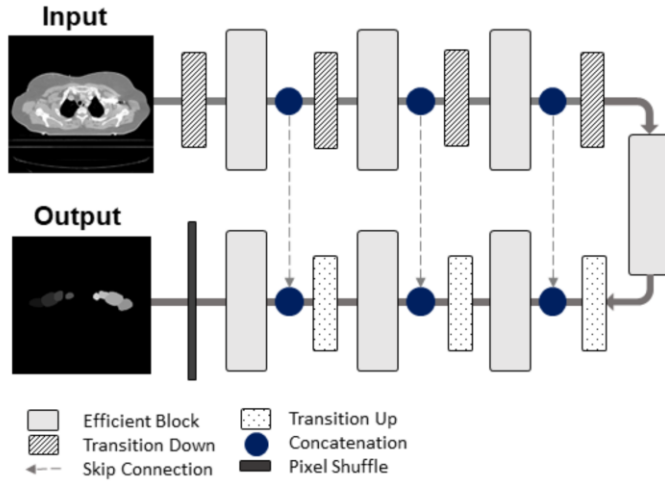


Fig 2. Schematic of the proposed convolutional neural network architecture (U-Net with EfficientNet-B0)

Among 111 cases that were newly contoured by an expert, a total of 92 cases were used as training dataset and 19 cases were used as test dataset #1 (contrast: 10 cases, non-contrast: 9 cases) for the analysis of quantitative metrics. Test dataset #2 was prepared separately to analyze the efficacy of auto-segmented contours using real-world heterogeneous data. Dosimetric parameters were analyzed using different sets of CT scans with manual contours (previously used for patient treatment) delineated by various physicians and RT plans of breast cancer patients who received RT after surgery ($n = 42$).

Both quantitative metrics and qualitative scoring were used for analyzing test dataset #1. Quantitative metrics included the most used geometrical indices, such as DSC and HD, to compare the auto-segmented and manually delineated contours. DSC is a measure of overlap between two contours, from “0” to “1,” where “1” indicates a complete overlap. HD is the measure of distance between two contours, where 0 mm indicates a complete overlap. For qualitative scoring,

two panels—an expert breast cancer radiation oncologist panel ($n = 11$) and a non-expert panel that included residents and radiation oncologists whose specialty is not breast cancer ($n = 15$)—from 10 institutions answered the following questions after watching an example video on manual contouring and auto-segmentation contouring on a planning CT scan:

What score would you give for the differences between manually delineated contours and auto-segmentation contours? (Difference scores)

Answer: 0 (most different) to 10 (least different)

How much do you think auto-segmentation would assist you in real-world clinical practice? (Assistance scores)

Answer: 0 (not helpful) to 10 (very helpful)

To analyze test dataset #2, auto-segmented contours were generated in 42 patients' CT scans, and dose-volume histograms were analyzed using both auto-segmented contours and original manual contours. Furthermore, dosimetric analysis was performed by comparing the mean dose (Gy), D0.03cc (Gy), and V5Gy (cc) for heart; mean dose (Gy), V20Gy, (%), and V5Gy (%) for ipsilateral lung; mean dose for contralateral lung; D0.03cc (Gy) for esophagus; and D1cc (Gy) for spinal cord for the manual and auto-segmented contours.

Additionally, inter-user variability was assessed by analyzing the DSCs and HDs of contours delineated by three different radiation oncologists on a randomly selected CT scan of a breast cancer patient. Furthermore, contouring time was recorded for all three radiation oncologists to compare the time taken

for manual delineation with that for auto-segmentation. Although the differences between the auto-segmented contours and manual contours were assessed quantitatively, in the field of radiation oncology, there is no precise answer or gold standard for CTV and OAR contours. Thus, differences do exist between contours delineated by different radiation oncologists. Table 1 shows the inter-observer variability through DSC and 95% HD for OARs and CTVs delineated by three board-certified radiation oncologists for a randomly selected patient. For OAR, only the heart and lungs showed a DSC above 0.80, whereas the other organs showed DSCs lower than 0.80. For CTV, although breast CTV showed an acceptable mean DSC of 0.85, other CTVs such as CTVn_L1, L2, L3, CTVn_IMN, CTVn_L4, and CTVn_SCL showed poor results, with mean DSC ranging from 0.45 to 0.75. For this randomly selected case, the contouring times for the three radiation oncologists were 35, 40, and 42 min, respectively, whereas the time taken to obtain auto-segmented contours was less than 10 min, including the time taken for sending the CT scan to the server and receiving the auto-segmented contours.

Table 1. Inter-observer variability of manual contours of organs-at-risk and target volumes

	DSC	STD	95% HD (mm)	STD (mm)
Organs-at-risk				
Heart	0.91	0.01	13.00	5.10
Rt Lung	0.99	0.00	2.33	0.95
Lt Lung	0.98	0.00	2.19	0.66
Thyroid	0.72	0.07	5.37	1.70
Esophagus	0.78	0.04	7.08	3.52
Spinal cord	0.69	0.09	72.89	49.91
Target				
CTVp_breast	0.85	0.02	8.94	2.86
CTVn_L1	0.69	0.04	13.58	3.00
CTVn_L2	0.47	0.17	18.74	8.15
CTVn_L3	0.56	0.10	9.87	3.61
CTVn_IMN	0.53	0.09	35.11	17.46
CTVn_L4	0.45	0.13	11.82	4.88
CTVn_SCL RTOG	0.75	0.03	6.93	0.62

3. EXTERNAL VALIDATION

Eleven experts who have a median of 7 (range, 2–21) years' experience of breast cancer radiotherapy was volunteered to participated in this study. The experts are attendings (n=2), clinical fellows (n=6), residents (n=2), and a dosimetrist (n=1) from two institutions (Yonsei Cancer Center and Asan Medical Center). Firstly, the 11 experts were requested to manually delineate OARs of breast cancer radiotherapy on simulation CT scans of 10 women planning to undergo radiotherapy for breast cancer (manual contours). The 9 OARs were thyroid, right lung, left lung, spinal cord, esophagus, heart, liver,

right breast, and left breast. Secondly, auto-contouring system was conducted for the same simulation CT scans and these auto-contours were provided to the experts. The experts were asked to correct the auto-contours as needed (corrected-auto-contours). Before contouring, the CT scans were de-identified, and the clinical information of patients was blinded. The clinical treatment contours that were used for the patient's radiotherapy delivery were removed to avoid bias during contouring. The experts were asked to record the video during the contouring for each CT scan using screen-recording software (oCam, OHSOFT, Korea). Additionally, as an exploratory analysis, a medical student with no experience in breast cancer radiotherapy performed the same procedure. This student's contouring metric data were shown in Fig 11 and not included in any other analyses.

Then, the best manual contours for each simulation CT were selected as ground truth after blind review by an independent third-party committee which consisted of five attending radiation oncologists who regularly treat breast cancer and have experience in breast cancer radiotherapy for more than 10 years. This review of contours was conducted online using a Google questionnaire platform without information on who contoured. Using these ground truths, accuracy was compared between the manual, corrected-auto, and auto-contour groups.

Each manual, corrected-auto-, and auto-contour was compared to the best manual contour using DSC and HD. Next, the sensitivity analysis was conducted; each contour was compared to the second-best manual contour, instead of the first-best manual contour, using DSC and HD. Whether the results achieved with the second-best manual contour are consistent with the primary results achieved with the first-best manual contour were assessed. (2) To assess

the time-saving effect, recorded videos on contouring were centrally reviewed and contouring times for all 9 OARs and for each OAR were measured. The time for manual contouring and correcting the auto-contours were compared. (3) To evaluate user satisfaction, simple questionnaires were sent to 11 experts to estimate the efficacy and feasibility of using auto-contouring system.

DSC and HD values were compared between the groups (manual, corrected-auto, and auto) using paired t-test. Contouring time was compared using Wilcoxon signed-rank test. P-values were corrected with Bonferroni's correction for group-wise comparisons. P-values <0.05 were considered statistically significant. Statistical calculations were performed using SPSS software (version 25; IBM Corp, Armonk, NY) and GraphPad Prism Version 8 (GraphPad Software Inc., La Jolla, USA).

III. RESULTS

1. COMPARISON with ABAS

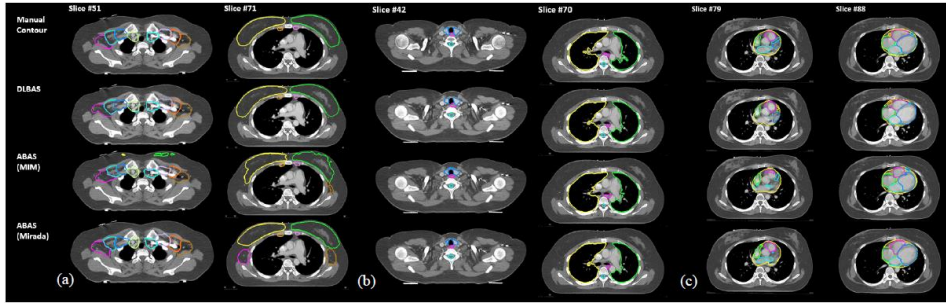


Fig 3. Examples of a) CTV, b) OAR, and c) heart segmentation results of DLBAS based on FCDN and ABAS by MIM and Mirada compared against ground-truth manual contours.

Figure 3A shows an example of CTV auto-segmentation from ABAS and DLBAS. Among 14 CTV structures, DLBAS produced the highest average DSCs in 11 of them. The statistical test reflects that these differences were significant for left and right AXL3 and IMN (Table 2). However, the HD comparisons of CTVs reveal that DLBAS produced smaller surface discrepancies compared to ABAS methods in every CTVs except for the SCL nodes as shown in Figure 4A. The difference was significant across most structures, except for the right AXL1 and SCL nodes.

As for the OARs, the performance of ABAS and DLBAS was comparable: an example is shown in Figure 4B. The highest average DSC and lowest HD for the lungs and spinal cord was produced by Mirada's ABAS (0.98 and 2.30mm), with the difference in the lungs being statistically significant (Table 3). The boxplots in Figure 4B show that DLBAS had quite large inter-subject variations

in the left lung and spinal cord. DLBAS did perform the best for thyroid and esophagus but not to a significant extent, as outlined in Table 3.

Figure 3C shows an example of heart auto-segmentation from ABAS and DLBAS. In the heart structures, DLBAS produced the highest average DSC in five out of seven heart structures (Table 4), with significantly higher results in the heart and right ventricle compared to both ABAS solutions. The HD comparison between ABAS and DLBAS describes that the average surface distance was significantly lower for DLBAS. This is further backed up by the smallest range of inter-subject variations in Figure 4C. The segmentation of the arteries (RCA and LAD) was below acceptable standards, with all solutions having less than a 50% match with the expert manual segmentation. Mirada's ABAS was unable to contour RCA for most test patients, so the HD results were excluded.

Table 2. Comparison of average DSC, HD and their significance for CTV segmentation of the patients in the test set (Mean + SD).

Right CTVs															Left CTVs									
Structure		Breast	AXL 1	AXL 2	AXL 3	IMN	SCL	SCL	Breast	AXL 1	AXL 2	AXL 3	IMN	SCL (E)	SCL (R)									
															(E)	(R)								
	Vol (ml)	607.0±	71.7±	51.9 ±	21.3 ±	24.6 ±	18.4 ±	48.9±	584.7±	72.6 ±	51.5±	22.0 ±	24.6 ±	20.5 ±	44.2 ±									
		165.6	18.7	12.7	5.8	3.9	5.3	10.1	119.7	21.1	12.6	6.8	3.7	10.7	11.0									
DSC	MIM	0.82 ±	0.75±	0.77±	0.67±	0.58±	0.67±	0.73±	0.83±	0.76±	0.72±	0.57±	0.59±	0.63±	0.74 ±									
		0.06	0.06	0.05	0.06	0.11	0.15	0.16	0.04	0.06	0.07	0.16	0.09	0.14	0.13									
	Mirada	0.89 ±	0.78±	0.79±	0.72±	0.68±	0.67±	0.80±	0.90±	0.76±	0.78±	0.66±	0.59±	0.65±	0.76 ±									
		0.03	0.04	0.06	0.04	0.08	0.12	0.03	0.02	0.07	0.06	0.08	0.08	0.12	0.12									
	FCDN	0.90 ±	0.75±	0.83±	0.79±	0.79±	0.77±	0.79±	0.90±	0.79±	0.80±	0.78±	0.72±	0.70±	0.75 ±									
		0.05	0.07	0.06	0.05	0.05	0.14	0.10	0.03	0.07	0.04	0.06	0.06	0.17	0.17									
Sig.	MIM vs Mirada	<0.01*	0.05	0.15	0.01*	<0.01*	0.99	0.12	<0.01*	0.76	0.01*	0.18	0.99	0.70	0.16									
	FCDN vs MIM	<0.01*	0.95	<0.01*	<0.01*	<0.01*	0.05	0.30	<0.01*	0.04	<0.01*	<0.01*	<0.01*	0.24	0.69									
	FCDN vs Mirada	0.60	0.18	0.02	<0.01*	<0.01*	0.01*	0.68	0.32	0.21	0.14	<0.01*	<0.01*	0.34	0.59									
HD	MIM	19.3 ±	11.2 ±	8.2 ±	8.1±	11.4 ±	15.3±	12.7 ±	15.2 ±	12.4	10.6	10.3	11.2	11.0 ±	9.8 ±									
		12.2	3.1	2.2	2.8	6.1	17.6	16.0	5.1	± 5.3	± 3.8	± 3.7	± 4.2	11.4	11.2									
	Mirada	9.4±3.4	10.3±	8.3±	6.8 ±	7.0±	10.6	6.6 ±	8.5±	11.2 ±	9.7 ±	7.7 ±	9.0±	10.3 ±	9.7 ±									
			4.3	2.9	2.0	2.4	± 10.4	1.8	3.2	4.9	5.2	2.3	2.7	10.7	10.8									

	FCDN		4.7 ± 1.7	8.7 ± 12.1	3.1 ± 1.1	2.7 ± 0.8	2.9 ± 1.1	3.8 ± 2.8	4.2 ± 1.8	4.3 ± 1.7	4.0 ± 1.3	3.4 ± 0.9	3.0 ± 1.2	3.7 ± 1.1	6.5 ± 5.2	8.6 ± 5.7
Sig.	MIM vs Mirada		<0.01*	0.50	0.89	0.06	<0.01*	0.27	0.19	<0.01*	0.61	0.48	0.06	0.10	0.44	0.81
	FCDN vs MIM		<0.01*	0.44	<0.01*	<0.01*	<0.01*	0.03	0.07	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	0.07	0.68
	FCDN vs Mirada		<0.01*	0.58	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	0.11	0.71

Table 3. Comparison of average DSC, HD and their significance for OAR segmentation of the patients in the test set (Mean + SD).

		Lung R	Lung L	Thyroid	Spinal Cord	Esophagus
Volume (ml)		1609.1 ± 399.6	1254.4 ± 385.2	25.7 ± 10.2	36.21 ± 67.0	52.3 ± 9.1
DSC	MIM	0.97 ± 0.01	0.97 ± 0.01	0.76 ± 0.07	0.79 ± 0.07	0.69 ± 0.10
	Mirada	0.98 ± 0.01	0.97 ± 0.01	0.79 ± 0.06	0.84 ± 0.04	0.72 ± 0.08
	FCDN	0.96 ± 0.01	0.95 ± 0.02	0.81 ± 0.06	0.81 ± 0.04	0.75 ± 0.05
Sig.	MIM vs Mirada	0.12	0.25	0.15	0.01*	0.05
	FCDN vs MIM	<0.01*	<0.01*	0.04	0.42	0.01*
	FCDN vs Mirada	<0.01*	<0.01*	0.29	0.10	0.28
HD	MIM	3.6 ± 2.8	2.8 ± 0.6	5.2 ± 2.3	3.5 ± 1.9	9.1 ± 4.8
	Mirada	2.3 ± 1.2	2.3 ± 0.9	3.9 ± 1.1	3.8 ± 3.3	7.0 ± 2.6
	FCDN	3.9 ± 1.9	6.5 ± 4.3	3.2 ± 1.6	6.4 ± 4.3	5.1 ± 1.9
Sig.	MIM vs Mirada	0.17	0.06	0.02	0.58	0.09
	FCDN vs MIM	0.67	<0.01*	<0.01*	0.06	<0.01*
	FCDN vs Mirada	<0.01*	<0.01*	0.13	0.18	0.04

Table 4. Comparison of average DSC, HD, and their significance for heart segmentation of the patients in the test set (Mean + SD).

		Heart	Atrium R	Atrium L	Ventricle R	Ventricle L	RCA	LAD
	Volume	726.0 ± 132.2	88.2 ± 18.4	73.4 ± 22.3	150.6 ± 28.9	199.4 ± 40.6	3.6 ± 1.7	6.5 ± 2.9
DSC	MIM	0.91 ± 0.03	0.79 ± 0.05	0.76 ± 0.08	0.78 ± 0.07	0.84 ± 0.05	0.10 ± 0.10	0.24 ± 0.15
	Mirada	0.94 ± 0.01	0.83 ± 0.03	0.82 ± 0.05	0.81 ± 0.038	0.88 ± 0.03	0.02 ± 0.06	0.04 ± 0.04
	FCDN	0.95 ± 0.01	0.85 ± 0.04	0.78 ± 0.04	0.86 ± 0.04	0.87 ± 0.03	0.32 ± 0.14	0.38 ± 0.15
Sig.	MIM vs Mirada	0.01*	<0.01*	0.04	0.11	0.03	<0.01*	<0.01*
	FCDN vs MIM	<0.01*	<0.01*	0.49	<0.01*	0.06	<0.01*	0.02
	FCDN vs Mirada	0.01*	0.09	0.03	<0.01*	0.10	<0.01*	<0.01*
HD	MIM	8.94 ± 4.36	9.79 ± 3.54	9.40 ± 4.26	10.68 ± 3.33	9.14 ± 3.21	20.06 ± 7.48	16.98 ± 7.93
	Mirada	6.09 ± 2.12	7.82 ± 2.22	8.57 ± 4.82	8.45 ± 2.06	6.95 ± 1.60	NA	52.00 ± 23.83
	FCDN	2.39 ± 0.47	4.03 ± 1.20	4.77 ± 2.18	6.15 ± 2.39	4.52 ± 1.65	8.40 ± 7.95	10.17 ± 6.78
Sig.	MIM vs Mirada	0.02	0.04	0.38	0.06	0.02	NA	<0.01*
	FCDN vs MIM	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*
	FCDN vs Mirada	<0.01*	<0.01*	<0.01*	<0.01*	<0.01*	NA	<0.01*

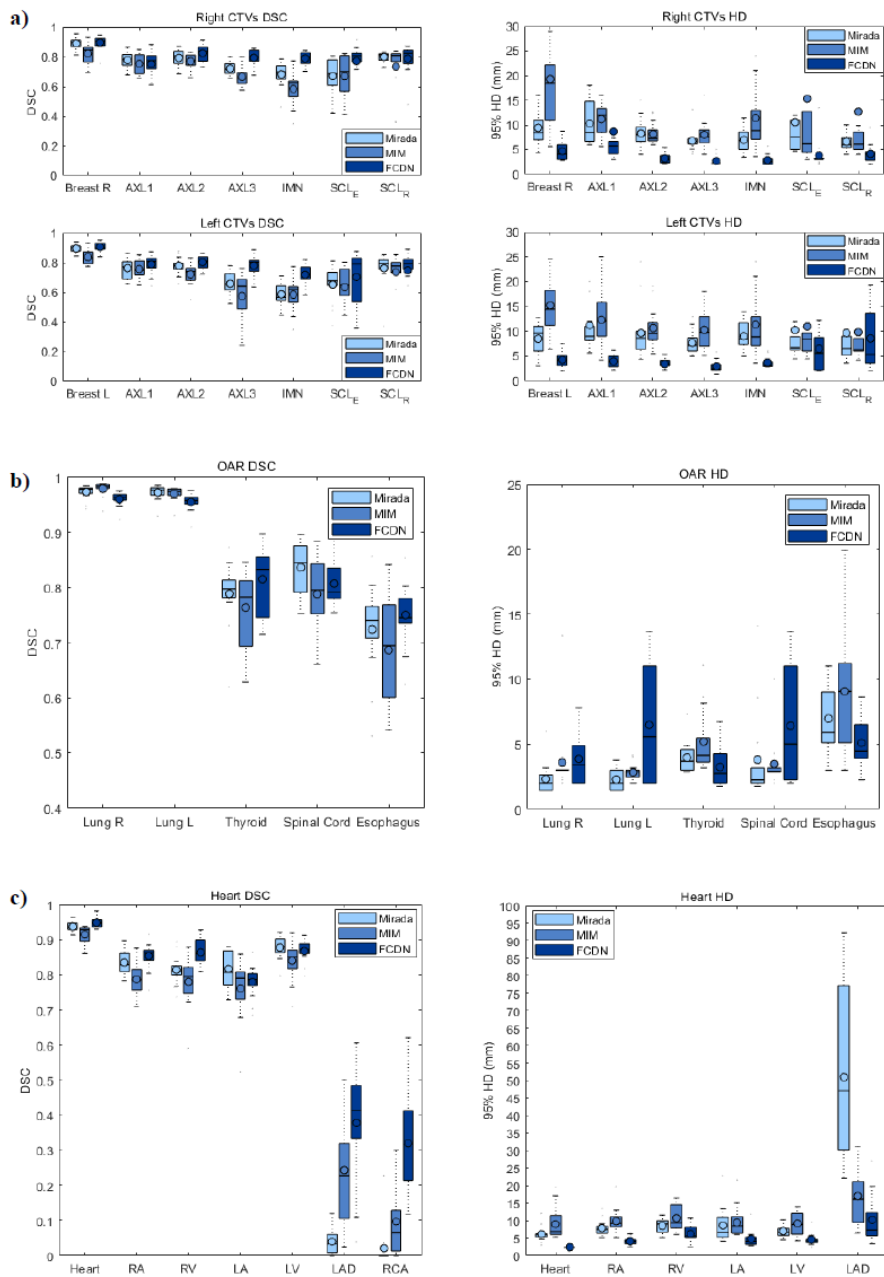


Fig 4. Box-plots of Dice Similarity Coefficients (DSC) and 95% Hausdorff Distance (HD) in the a) CTVs, b) OARs, and c) Heart structures obtained from

Mirada, MIM, and DLBAS based on FCDN using the manual contours as reference.

Lastly, Figure 5 shows the results of sensitivity analysis DLBAS and ABAS, where the bars indicate the Δ DSC, defined as DSCs of non-contrast test data subtracted from DSCs of contrast test data. DLBAS showed much smaller Δ DSC compared to ABAS, especially for the CTVs. Similarly, DLBAS produced the lowest difference between HD values of contrast and non-contrast test set (Figure 6).

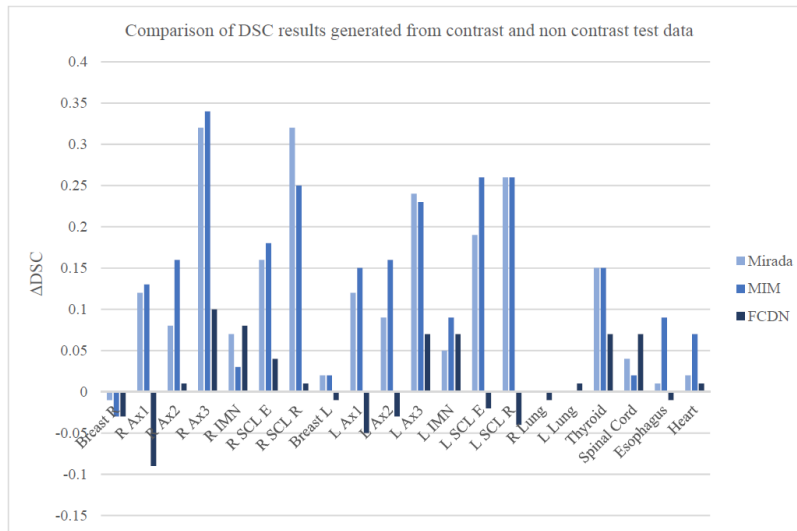


Fig 5. Difference in dice similarity coefficients (Δ DSC) between contrast and non-contrast test sets obtained from DLBAS based on FCDN and ABAS.

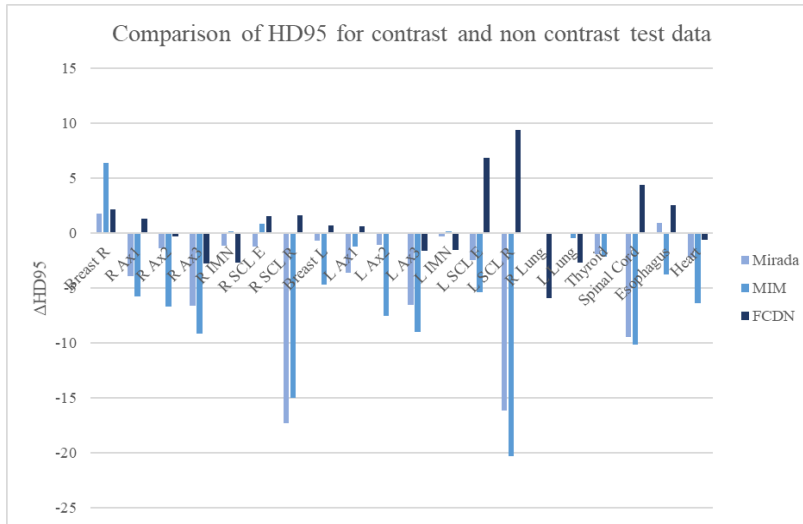


Fig 6. Difference in 95% Hausdorff distance (Δ HD) between contrast and non-contrast test sets obtained from fully convolutional DenseNet (FCDN) and atlas-based auto-segmentation (ABAS).

2. INVESTIGATION OF CLINICAL USEFULNESS

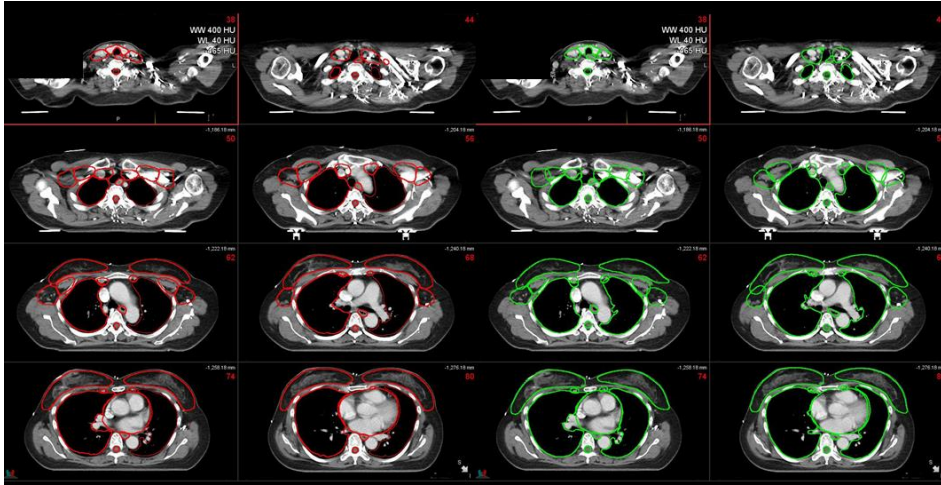


Fig 7. Example of deep learning-based auto-segmentation (green) and manual contours (red).

Examples of DLBAS and manual contours are shown in Figure 7. Table 5 compares the auto-segmented contours and manual contours for OARs and CTVs using mean DSC and 95% HD. Regarding OARs, mean DSCs were above 0.80 and mean 95% HDs were below 5 mm, which are acceptable results. For CTV, the correlation between the auto-segmented and manual contours was excellent for breast, with a mean DSC higher than 0.90. As for other CTVs, including CTVn_L1, L2, L3, CTVn_IMN, CTVn_L4, and CTVn_SCL RTOG, the mean DSCs were mostly higher than 0.70. The mean 95% HD ranged from 5.50 to 10.93 mm for CTVs. The mean DSCs and 95% HDs did not show a large difference between the contrast-enhanced CT test datasets and non-contrast CT test datasets.

Figures 8A and 8B show dose-volume histograms with average dosimetric values for patients who received whole breast RT only or that with

regional node irradiation, respectively. The increase at the end for the ipsilateral breast contour line in the dose-volume histograms is due to the initial RT plan that included a simultaneous integrated boost for the tumor bed. As shown in Figure 8A, most manual and auto-segmented contours were similar, except for a minor difference for the spinal cord. The difference in the delineated spinal cord volume (average absolute difference of 7.24 ± 9.07 cc) may have affected the results. Figure 8B shows that there was a considerable difference in the coverage for regional nodal contours such as axillary lymph node levels 1, 2, 3, and IMN.

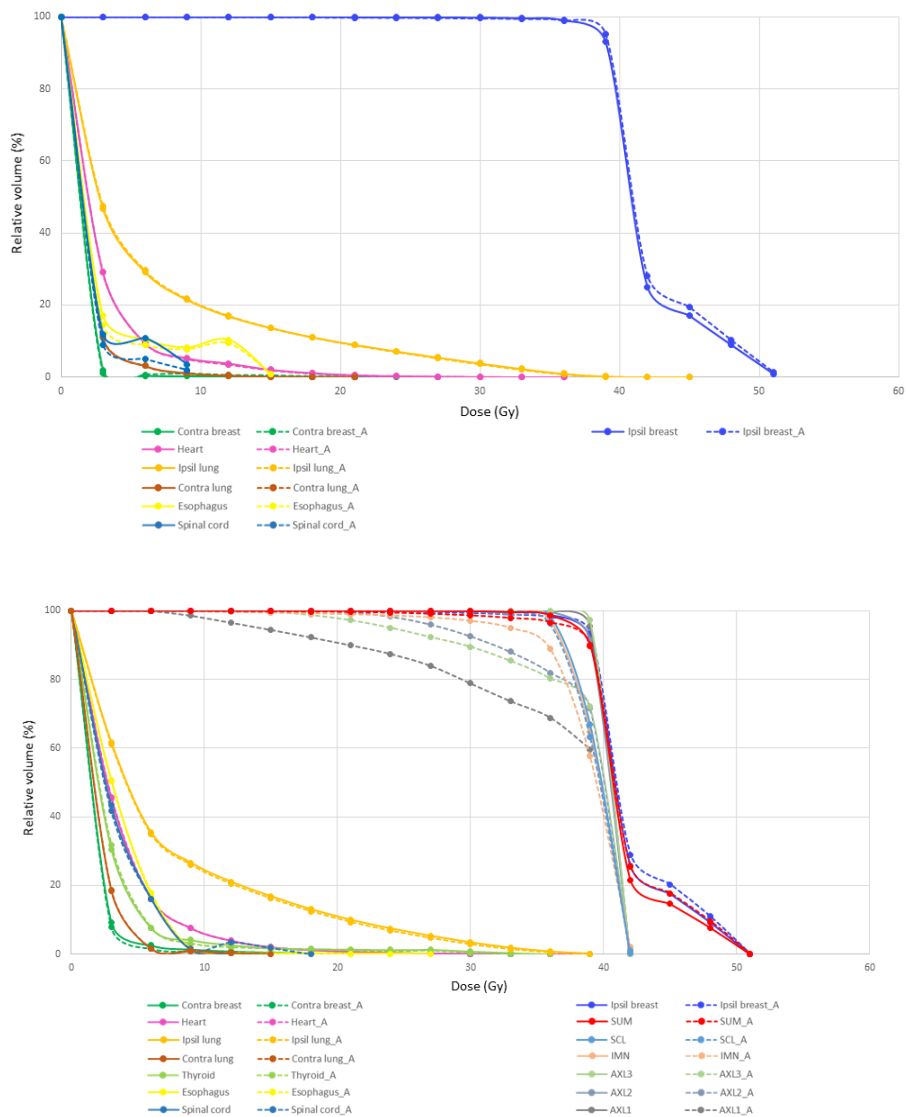


Fig 8. Comparison of dose-volume histograms with average dosimetric values of manual contours (solid line) and auto-segmentation contours (dotted line) for patients who received whole breast RT only (A) or that with regional node irradiation (B).

In addition, various dosimetric parameters for OARs—such as heart, lung, esophagus, and spinal cord—were analyzed, as shown in Table 6. The mean absolute differences for all parameters were minimal, showing the efficacy of auto-segmented contours.

Table 5. Comparison of deep learning auto-segmentation and manual contours of organs-at-risk and target volumes

	Total (n=19)				Contrast (n=10)				Non-contrast (n=9)			
	DSC	STD	95% HD (mm)	STD (mm)	DSC	STD	95% HD (mm)	STD (mm)	DSC	STD	95% HD (mm)	STD (mm)
Organs-at-risk												
Heart	0.95	0.02	4.56	2.33	0.96	0.01	3.83	2.80	0.94	0.02	5.36	1.27
Rt Lung	0.98	0.01	3.61	2.15	0.98	0.00	4.64	2.46	0.97	0.01	2.46	0.69
Lt Lung	0.97	0.01	2.82	0.71	0.97	0.01	3.04	0.76	0.97	0.02	2.59	0.55
Thyroid	0.89	0.05	1.88	0.90	0.90	0.04	1.55	0.65	0.88	0.05	2.25	0.99
Esophagus	0.84	0.06	2.87	1.49	0.85	0.05	2.47	0.91	0.83	0.07	3.31	1.85
Spinal cord	0.82	0.10	2.98	3.10	0.87	0.07	1.58	0.74	0.76	0.10	4.54	3.89
Target												
CTVp_breast	0.94	0.04	5.50	3.17	0.94	0.04	5.13	2.74	0.94	0.04	5.91	3.55
CTVn_L1	0.74	0.08	10.93	6.27	0.71	0.09	13.51	7.10	0.78	0.05	8.07	3.40
CTVn_L2	0.80	0.07	6.36	2.52	0.79	0.07	6.71	2.40	0.81	0.06	5.98	2.60
CTVn_L3	0.64	0.13	7.99	3.81	0.66	0.10	6.97	2.87	0.62	0.16	9.11	4.37
CTVn_IMN	0.72	0.09	5.75	3.36	0.67	0.09	7.53	3.71	0.77	0.07	3.77	1.00
CTVn_L4	0.74	0.12	6.04	6.12	0.67	0.12	8.37	7.61	0.80	0.09	3.45	1.41
CTVn_SCL	0.78	0.08	6.95	2.89	0.76	0.08	7.85	3.20	0.80	0.08	5.95	2.09

RTOG

Table 6. Dosimetric outcomes for manual and auto-segmented contours.

	Manual		Autocontour		Absolute difference	
	Mean	STD	Mean	STD	Mean	STD
Heart						
Mean (Gy)	3.27	1.10	3.26	1.10	0.08	0.07
D _{0.03cc} (Gy)	22.72	10.24	21.75	9.68	1.51	1.76
V _{5Gy} (cc)	16.08	10.23	16.13	10.39	0.73	0.75
Lung						
Ipsilateral lung mean (Gy)	6.87	0.97	6.82	0.97	0.11	0.18
Ipsilateral lung V _{20Gy} (%)	10.08	2.59	9.82	2.68	0.36	0.35
Ipsilateral lung V _{5Gy} (%)	35.25	5.15	35.46	5.32	0.67	0.98
Contralateral lung mean (Gy)	1.83	0.66	1.84	0.67	0.02	0.03
Esophagus						
D _{0.03cc} (Gy)	7.82	5.16	7.56	4.73	0.85	1.61
Spinal cord						
D _{1cc} (Gy)	4.38	2.69	4.37	2.96	0.43	1.08

To confirm whether DLBAS can practically serve as a useful tool in clinical practice, qualitative scores were also analyzed. Qualitative scoring was performed by both an expert ($n = 11$) and a non-expert panel ($n = 15$) for difference and assistance scores, as shown in Figure 4. For OARs, the median difference score was 9 (range, 8–10) and the median assistance score was 9 (range, 8–10), in the case of the expert panel. The scores were similar for OARs in the case of the non-expert panel, with a median difference score of 8 (range, 6–10) and a median assistance score of 9 (range, 8–10). For CTVs of breasts and regional lymph nodes, the median difference score was 8 (range, 7–9) and the median assistance score was 9 (range, 7–10), in the case of the expert panel. Regarding the non-expert panel, the median difference score was 8 (range, 6–10) and the median assistance score was 9 (range, 5–10).

3. EXTERNAL VALIDATION

After conducting the process of the study, we collected 110 manual contours, 110 corrected-auto-contours, and 10 auto-contours of each type of OAR. When these contours were compared to the consensus ground truth contours, 100 DSC and 100 HD values (pairs of the ground truth contour and each contour) were created for each types of OARs for manual and corrected-auto-contours, respectively, and 10 DSC and 10 HD values for auto-contours.

Table 7 and Figure 9 shows the mean DSC and HD values of manual, corrected-auto-, and auto-contours by each OAR. The DSC of breast contours were prominently higher in the corrected-auto-contours or auto-contours than in the manual contours; the absolute differences of DSCs between the manual and corrected-auto contours were 0.09 in the right breast and 0.07 in the left breast, while those of the other OARs were less than 0.03. The HD values of the breast, heart, and liver contours showed prominent differences; the HDs of breast and heart contours were prominently lower in the corrected-auto-contours or auto-contours compared than in the manual contours. In contrast, the HD of liver contours was prominently higher in the auto-contours than in the manual or corrected-auto-contours.

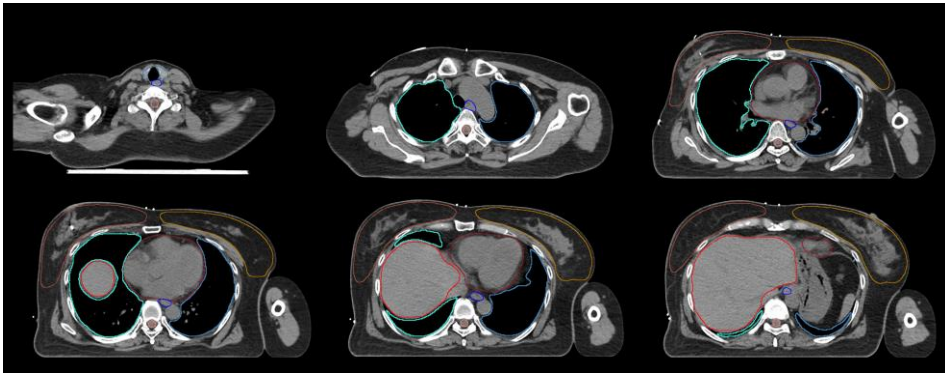


Fig 9. Examples of deep learning-based auto-contour without correction.

In the cases of the breast, spinal cord, and heart contours, corrected-auto-contours had better accuracy than manual contours, which is consistently seen in DSC and HD values. In contrast, in cases of the thyroid and lung contours, the manual contour had better accuracy than the corrected-auto-contour, which is consistently seen in DSC and HD values. In the cases of liver and esophagus contours, mixed results were shown; DSCs were higher in the manual contours, but the HDs were lower in the corrected-auto-contours. The results of the sensitivity analyses were in consistent with the original analyses (Table 8 and Fig 10). The results of statistical analyses for contour comparisons are shown in Table 8 and 9.

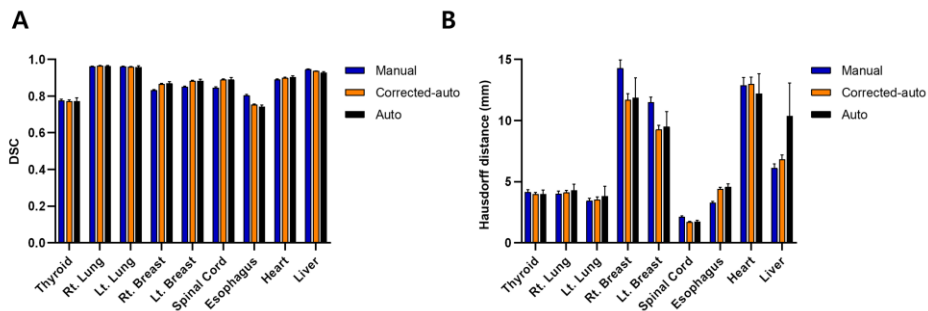


Fig 10. Dice similarity coefficient (A) and Hausdorff distance values (B) according to the organ-at-risks, comparing manual contours, corrected-auto-contours, and auto-contours. For sensitivity analyses, contouring metrics were obtained by comparing each contour with the second-best contour. Data are shown in mean \pm standard error.

To evaluate the performance of auto-contour itself, the DSC and HD values of all OARs were compared between manual and auto-contours. In manual contours, the mean DSCs of all OARs ranged from 0.86 to 0.90 (median, 0.88) according to the experts, and HDs of all OARs ranged from 5.14 to 9.09 mm

(median, 6.44 mm). Based on the mean DSCs of all OARs, the auto-contour ranked 2nd place with a value of 0.896, following the expert whose value was 0.903. Based on the mean HDs of all OARs, auto-contour ranked 1st place with a value of 5.142 mm, followed by the expert whose value was 5.327 mm (Table 10).

The inter-physician variations observed in the experts' manual contours were reduced in the corrected-auto-contours. The range of mean DSC of all OARs was 0.86–0.90 according to the individuals in manual contours but reduced to 0.89–0.90 in corrected-auto-contours. The range of mean HD of all OARs was 5.14–9.19 mm according to the individuals in manual contours but reduced to 4.3–5.7 mm in corrected-auto-contours. Figure 11 shows the mean DSCs according to the OARs. The figure shows that DSCs were more homogeneous in the corrected-auto-contours than in the manual contours, meaning that inter-physician variability was reduced. Contours of a medical student were also shown in the figure, showing the improved accuracy in the corrected-auto-contours like other experts. A sensitivity analysis was shown in figure 12 and showed the consistent results to the original analyses. Examples of manual and corrected-auto contours of breast and heart are shown in Fig 13. Notably, the inter-physician variability of manual breast contours was mostly seen in the lateral and anterior borders of the breast, while this variability was rarely seen in corrected-auto-contours.

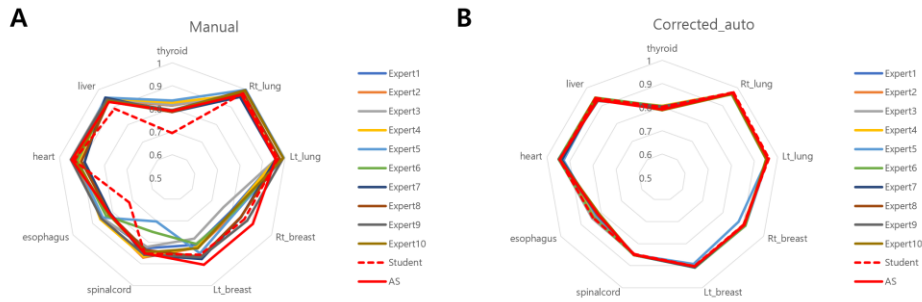


Fig 11. Radar graphs showing mean DSC value of each participant according to the organs: (A) manual contours, (B) corrected-auto-contours. DSC values of corrected-auto-contours were more homogeneous than those of manual contours meaning that inter-physician variability was reduced.

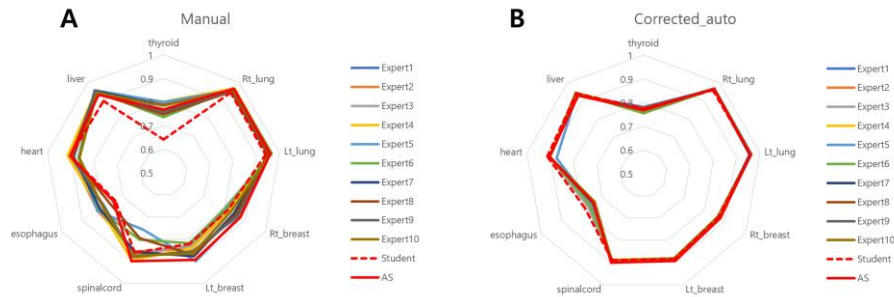


Fig 12. Radar graphs showing mean DSC value of each participant according to the organs: (A) manual contours, (B) corrected-auto-contours. DSC values of corrected-auto-contours were more homogeneous than those of manual contours meaning that inter-physician variability was reduced. For sensitivity analyses, contouring metrics were obtained by comparing each contour with the second-best contour.

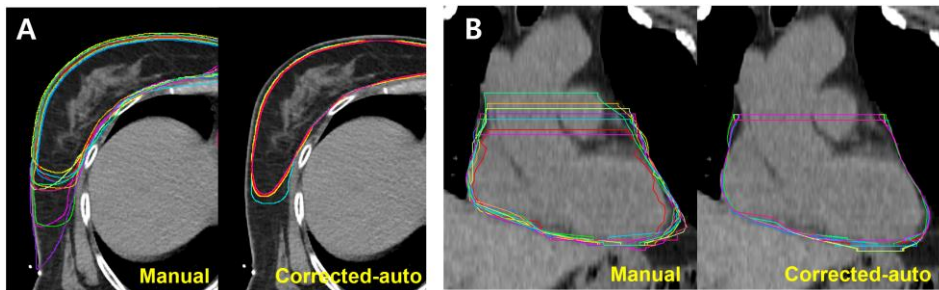


Fig 13. Examples of manual and corrected-auto-contours of all experts: (A) breast contours showing that inter-physician variability is mostly seen in lateral and anterior borders of breasts and is reduced with an aid of auto-contouring system, as well as (B) heart contours.

Mean contouring time for 9 OARs of each patient was 37 min (standard deviation [SD], 20 min) for manual contours and 6 min (SD, 5 min) for corrected-auto-contours, showing 84%-time reduction with an aid of auto-contouring system (Fig. 14A and Table 11). When mean time was measured according to each OAR, breast and liver contouring was the longest step among the manual contours. The time was prominently reduced in the corrected-auto-contours (right breast: 5.9 min [SD, 3.8 min] to 0.5 min [SD, 1.1 min]; left breast: 6.3 min [SD, 4.1 min] to 0.6 min [0.7 min]; liver: 9.0 min [SD, 5.0 min] to 1.5 min [SD, 1.2 min]) (Fig. 14B and Table 12 and 13).

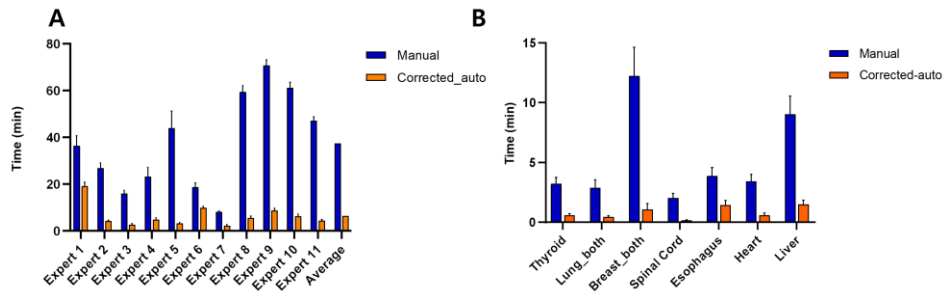


Fig 14. Contouring time comparing manual contouring and correcting auto-contours: (A) total contouring time of all 9 organ-at-risks of each expert, (B) contouring time of each organ-at-risks. Data are shown in mean \pm standard error.

The mean score of the question, “how was the accuracy of auto-contours?”, was 7.5 ± 0.9 , that of the question, “how much auto-contours helped to shorten the contouring time?”, was 8.8 ± 1.1 , and that of the question, “Do you want to use auto-contours in future practice”, was 9.2 ± 0.7 , where the answers were numerical values ranging 0 (worst) to 10 (best).

Table 7. Summary of Dice Similarity Coefficient and Hausdorff distance

	Dice Similarity Coefficient (mean±SD)			Hausdorff distance (mean±SD)		
	Manual contour	Corrected-auto-contour	Auto-contour	Manual contour	Corrected-auto-contour	Auto-contour
Thyroid	0.8±0.06	0.8±0.06	0.79±0.07	3.82±2.03	4.12±2.56	4.28±2.62
Lung_right	0.98±0.02	0.97±0.01	0.97±0.01	2.37±1.49	2.46±0.92	2.42±0.93
Lung_left	0.97±0.02	0.96±0.02	0.96±0.02	3.61±3.46	2.93±1.77	2.99±2.11
Breast_right	0.81±0.05	0.9±0.02	0.91±0.01	12.44±7.96	8.06±3.48	7.54±2.05
Breast_left	0.83±0.04	0.9±0.02	0.9±0.03	10.85±5.83	8.14±3.29	8.15±3.32
Spinal cord	0.82±0.07	0.85±0.03	0.85±0.03	3.95±5.37	2.2±0.35	2.21±0.36
Esophagus	0.84±0.04	0.83±0.03	0.82±0.03	3.95±3.64	3.16±0.59	3.46±0.68
Heart	0.92±0.03	0.94±0.01	0.95±0.01	8.26±5.47	5.42±2.85	4.73±1.05
Liver	0.94±0.02	0.94±0.02	0.93±0.02	6.92±6.36	5.95±3.22	9.74±9.98

Abbreviation: SD, standard deviation; *p*-values can be found in Supplementary files

Table 8. Summary of Dice Similarity Coefficient and Hausdorff distance

		Contour			<i>P</i> -value*		
		(1) Manual	(2) Corrected-auto	(3) Auto	(1) vs (2)	(1) vs (3)	(2) vs (3)
DSC (mean±SD)	Thyroid	0.8±0.06	0.8±0.06	0.79±0.07	.953	.042	.014
	Lung_right	0.98±0.02	0.97±0.01	0.97±0.01	<.001	<.001	.262
	Lung_left	0.97±0.02	0.96±0.02	0.96±0.02	<.001	<.001	.801
	Breast_right	0.81±0.05	0.9±0.02	0.91±0.01	<.001	<.001	.596
	Breast_left	0.83±0.04	0.9±0.02	0.9±0.03	<.001	<.001	.280
	Spinal cord	0.82±0.07	0.85±0.03	0.85±0.03	.001	.001	.136
	Esophagus	0.84±0.04	0.83±0.03	0.82±0.03	.001	<.001	<.001
	Heart	0.92±0.03	0.94±0.01	0.95±0.01	<.001	<.001	.029
HD (mean±SD)	Liver	0.94±0.02	0.94±0.02	0.93±0.02	<.001	<.001	<.001
	Thyroid	3.82±2.03	4.12±2.56	4.28±2.62	0.33	0.056	0.021
	Lung_right	2.37±1.49	2.46±0.92	2.42±0.93	>.999	>.999	0.472
	Lung_left	3.61±3.46	2.93±1.77	2.99±2.11	0.077	0.115	0.538
	Breast_right	12.44±7.96	8.06±3.48	7.54±2.05	<.001	<.001	0.284
	Breast_left	10.85±5.83	8.14±3.29	8.15±3.32	<.001	<.001	>.999
	Spinal cord	3.95±5.37	2.2±0.35	2.21±0.36	0.005	0.005	0.475
	Esophagus	3.95±3.64	3.16±0.59	3.46±0.68	0.1	0.574	<.001

Heart	8.26±5.47	5.42±2.85	4.73±1.05	<.001	<.001	0.034
Liver	6.92±6.36	5.95±3.22	9.74±9.98	0.335	0.085	0.002

Abbreviations: DSC, Dice similarity coefficient; HD, Hausdorff distance; SD, standard deviation

* *P*-values were calculated using paired t-test with bon-ferroni correction.

Table 9. Summary of Dice Similarity Coefficient and Hausdorff distance for sensitivity analyses.

		Contour			<i>P</i> -value*		
		Manual	Corrected-auto	Auto	MS vs AS+R	MS vs AS	AS+R vs AS
DSC (mean±SD)	Thyroid	0.78±0.07	0.77±0.06	0.77±0.06	0.942	1.031	>.999
	Lung_right	0.96±0.02	0.97±0.01	0.96±0.01	0.099	0.214	0.336
	Lung_left	0.96±0.02	0.96±0.02	0.96±0.02	>.999	0.441	0.04
	Breast_right	0.83±0.04	0.87±0.03	0.87±0.03	<.001	<.001	0.355
	Breast_left	0.85±0.04	0.88±0.03	0.88±0.03	<.001	<.001	0.971
	Spinal cord	0.85±0.06	0.89±0.03	0.89±0.03	<.001	<.001	>.999
	Esophagus	0.8±0.04	0.75±0.03	0.74±0.03	<.001	<.001	<.001
	Heart	0.89±0.03	0.9±0.03	0.9±0.02	0.007	0.001	0.066
	Liver	0.95±0.02	0.94±0.02	0.93±0.02	<.001	<.001	<.001
HD (mean±SD)	Thyroid	4.16±1.94	3.99±1.28	3.98±1.07	0.956	0.962	>.999
	Lung_right	4.04±2.01	4.12±1.62	4.29±1.63	>.999	0.421	0.003
	Lung_left	3.45±2.01	3.54±1.98	3.84±2.49	>.999	0.045	0.006
	Breast_right	14.28±6.64	11.7±4.88	11.87±5.11	0.001	0.005	>.999
	Breast_left	11.51±3.98	9.27±3.54	9.5±3.87	<.001	0.002	0.524
	Spinal cord	2.15±0.5	1.71±0.34	1.73±0.32	<.001	<.001	0.134
	Esophagus	3.29±1.04	4.4±1.27	4.6±0.76	<.001	<.001	0.411

Heart	12.9±6.36	12.99±5.59	12.21±5.11	>.999	0.99	0.015
Liver	6.11±3.24	6.83±3.46	10.37±8.53	0.258	<.001	0.001

Abbreviations: DSC, Dice similarity coefficient; HD, Hausdorff distance; SD, standard deviation

* *P*-values were calculated using paired t-test with bon-ferroni correction.

Table 10. The DSC and HD values of all OARs of experts' manual contours and an auto-contour, listed from the best to the lowest performance.

Rank	DSC	HD
	Average (Standard deviation)	
1	0.903 (0.065)	5.142 (2.692)*
2	0.896 (0.06)*	5.327 (3.623)
3	0.887 (0.062)	5.477 (3.055)
4	0.886 (0.061)	5.615 (3.543)
5	0.882 (0.066)	5.78 (3.72)
6	0.881 (0.06)	6.431 (4.818)
7	0.881 (0.076)	6.447 (5.938)
8	0.88 (0.09)	6.461 (4.021)
9	0.877 (0.067)	6.501 (4.072)
10	0.874 (0.08)	6.724 (6.298)
11	0.87 (0.087)	7.636 (7.014)

* The value of an auto-contour.

Table 11. Total contouring time for 9 organ-at-risks of each patient

	Manual (min)	Correced-auto (min)	Difference* (min)
Expert 1	36.4	19.1	17.3
Expert 2	26.9	4.2	22.6
Expert 3	16.0	2.6	13.4
Expert 4	23.2	4.8	18.4
Expert 5	43.9	3.2	40.7
Expert 6	18.7	9.8	8.9
Expert 7	8.1	2.2	5.9
Expert 8	59.4	5.5	53.9
Expert 9	70.8	8.7	62.1
Expert 10	61.2	6.2	55.1
Expert 11	47.1	4.4	42.7
Mean	37.4	6.4	31.0
SD	19.7	4.6	19.4

Abbreviations: SD, standard deviation

* Time for correcting auto-contours minus time for manual contouring

Table 12. Time for manual contouring according to the organ-at-risks

	Thyroid (min)	Lung_right (min)	Lung_left (min)	Breast_right (min)	Breast_left (min)	Spinal Cord (min)	Esophagus (min)	Heart (min)	Liver (min)
Expert 1	4.3	1.3	1.5	5.1	6.1	1.2	4.4	3.7	8.0
Expert 2	1.7	1.4	0.6	5.8	6.2	1.2	1.6	1.7	5.9
Expert 3	1.7	0.4	0.8	1.8	1.7	0.9	2.3	1.6	3.6
Expert 4	2.0	0.5	0.6	1.8	2.4	1.3	2.3	1.6	7.2
Expert 5	2.4	4.9	3.9	6.1	6.7	3.2	4.1	4.3	7.5
Expert 6	1.8	0.9	0.9	2.6	2.3	0.8	2.1	1.7	5.7
Expert 7	1.1	0.6	0.4	0.8	0.8	0.7	0.5	1.1	1.6
Expert 8	4.0	3.3	2.2	11.2	10.7	4.0	6.7	6.2	11.1
Expert 9	5.4	1.9	1.3	12.8	14.8	4.3	7.9	6.2	16.1
Expert 10	7.0	1.7	1.0	8.7	8.9	2.8	7.3	6.2	17.6
Expert 11	4.1	1.2	0.6	8.7	8.7	2.0	3.4	3.3	15.1
Mean	3.2	1.6	1.2	5.9	6.3	2.0	3.9	3.4	9.0
SD	1.8	1.3	1.0	3.8	4.1	1.3	2.4	2.0	5.0

Abbreviations: SD, standard deviation

Table 13. Time for correcting auto-contours according to the organ-at-risks

	Thyroid (min)	Lung_right (min)	Lung_left (min)	Breast_right (min)	Breast_left (min)	Spinal Cord (min)	Esophagus (min)	Heart (min)	Liver (min)
Expert 1	1.4	0.2	1.2	3.9	2.4	0.6	2.5	2.4	3.5
Expert 2	0.1	0.1	0.1	0.4	0.5	0.0	0.2	0.4	1.2
Expert 3	0.1	0.1	0.1	0.0	0.2	0.0	0.8	0.0	0.6
Expert 4	0.7	0.1	0.1	0.1	0.2	0.2	0.6	0.4	1.0
Expert 5	0.1	0.5	0.2	0.1	0.1	0.3	0.1	0.0	0.4
Expert 6	1.1	0.1	0.0	0.1	0.9	0.1	1.5	1.2	4.2
Expert 7	0.1	0.2	0.1	0.0	0.0	0.1	0.0	0.3	0.7
Expert 8	0.7	0.1	0.2	0.6	1.1	0.2	1.8	0.2	0.6
Expert 9	1.1	0.7	0.3	0.2	0.6	0.3	3.6	1.0	1.0
Expert 10	0.8	0.1	0.1	0.0	0.1	0.0	3.7	0.1	1.4
Expert 11	0.5	0.2	0.2	0.0	0.1	0.0	1.1	0.4	1.9
Mean	0.6	0.2	0.2	0.5	0.6	0.2	1.5	0.6	1.5
SD	0.4	0.2	0.3	1.1	0.7	0.2	1.3	0.7	1.2

Abbreviations: SD, standard deviation

IV. DISCUSSION

A. COMPARISON with ABAS

To the best of our knowledge, this is the first study that compares the performance of ABAS and DLBAS methods for breast cancer RT planning that includes node regions and heart structures. In this study, we demonstrated the efficacy of our DLBAS by measuring the performance across a range of structures (CTVs, OARs, and heart) in a head-to-head comparison study with commercial ABAS solutions. Our results indicate that, while ABAS offered an acceptable performance level, DLBAS showed much more robust and reliable automatic segmentation outcomes that were in greater spatial similarities with the ground-truth. These findings are parallel with several studies comparing ABAS and DLBAS methods in other malignancies. Lustberg et al. were the first to compare the delineation by a commercial DLBAS program with those obtained by an ABAS program in five OARs for lung cancer.³⁴ Other studies have compared the ability of a commercial ABAS program and a convolutional neural network (CNNs)-based DLBAS to delineate OARs found in liver cancer³⁵ and lung cancer; these studies provided evidence that DLBAS is more accurate and computationally faster than ABAS.

DLBAS had a competitive edge on their ABAS counterparts in general, especially in some CTVs (e.g. AXL3 and IMN) and heart structures in terms of both DSC and HD. Our finding with respect to high performance of DLBAS in heart structures (overall chamber accuracy of 0.86) is better, or at least similar, than previous studies. Jung et al. reported average DSCs of 0.66 in the atria and 0.75 in the ventricles with ABAS approach.³⁶ Dormer et al. reported good performance with a two-dimensional CNN-based segmentation method for

heart substructures trained on 11 3D CT data.³⁷ In this study, the performance of ABAS was below a satisfactory level in coronary arteries (RCA, LAD), which is similar with two previous studies^{37,38} that both reported a mean DSC less than 0.3. On the other hand, DLBAS and ABAS demonstrated a comparable performance in the OARs. The OAR structures tend to have distinct structural boundaries, which would have worked in favour of ABAS' intensity-based registration.

Our results are indicative of three important findings for DLBAS. Firstly, we highlighted the ability of DLBAS in learning the characteristics of complex and low-contrast anatomy, implying its potential for CTV delineation where the quality of delineation is highly dependent on expert knowledge. Conversely, it indicates the limitation of ABAS in this perspective as it is based on landmark-based detection. Secondly, the fact that DLBAS has much smaller contouring differences with the ground-truth (on the order of millimeters as indicated by HD) highlights the clinical impact of its implementation. Although widely used in the literature, DSC is often highly affected by the volume size of the two contours under comparison which is the reason why distance metrics like HD were introduced for evaluation. If auto-segmented contours were to be modified in a clinical setting, HD values are likely to be an important factor that determines the time taken for modifications as its values are directly linked to the accuracy of the contour outlines, more so than the DSC. Based on the statistical analysis of our DSC and HD results, we believe that DLBAS could be a solution that can reduce the time required for generating contours of the CTVs and heart substructures than ABAS. Thirdly, our finding with respect to robustness of DLBAS on non-contrast CT test samples is noteworthy. DLBAS showed much smaller DSC discrepancies between contrast and non-contrast results compared to ABAS. These findings further indicate that our DLBAS

model is less dependent on the type of input than ABAS, hence could be more robust, making it is beneficial in clinical settings where patient data are not always collected with consistent protocols such as CT contrast.

Based on the results, DLBAS has great potential to be the solution for many issues encountered in RT planning. First and foremost, it can address the consistency issue of manual contouring in practice and large-scale clinical studies. Weber et al. discussed that large inter-physician and inter-institutional variations are one of the biggest challenges in today's radiation oncology trials.³⁹ Up to 13.4% of assessed RT plans in clinical trials were deemed unacceptable due to major deviations in target volume delineations.⁴⁰ This is problematic due to the associated safety issues and toxicity from incorrect delineations. A recent survey that investigated inter-institutional variations in breast IMRT in South Korea has revealed that there are large heterogeneities in the target volume and RT plans between practices. If these contours can be generated in a consistent manner using DLBAS, the burden and possible complications in these trials may be avoided. However, implementing DLBAS comes with initial time and cost investments involving purchasing or building a reliable DLBAS model specifically designed for each institution, which involves patient data collection and generation of the ground-truth labels by the experts. However, once this foundation work is completed, the time spent for repetitive work (i.e. contouring) will be greatly reduced, which would subsequently allow us to attend to other clinical activities in the department.

B. INVESTIGATION OF CLINICAL USEFULNESS

Our findings suggest that the proposed algorithm performed well, exhibiting good agreement with the CTVs and OARs that were manually contoured by

clinical experts from both qualitative and quantitative aspects. The dosimetric implications of the auto-contours were also evaluated, and we did not observe any significant difference in dose-volume histograms between the auto-segmented contours and manual contours.

Although AI solutions are best suited to situations in radiology where ground truths are clear, the concept of a ground truth in RT fields is disputable because RT is both a science and an art entailing clinical input and creativity⁴¹. More specifically, inter-physician variations are present even in contours delineated by board-certified radiation oncologists from the same institution (e.g., variations in the nodal target volumes in our study; Additional file 1). We acknowledge that the generation of the same contours by an AI algorithm under multiple scenarios does not mean that the generated contours are optimal. Considering this, we collected data based on the assumption that the international guidelines are an alternative ground truth.^{28,42} Although the proposed algorithm performed well, a risk exists that its reliability may decrease in some situations.⁴³

In 2006, Eldesoky et al. first reported the clinical utility of ABAS in loco-regional RT for breast cancer using the data of 60 patients, where delineation was performed according to the ESTRO consensus guideline.⁴⁴ ABAS showed good agreement in some volumes (e.g., lung, heart, and breast), whereas it showed only modest agreement in other structures or in external datasets. However, research interest shifted to DLBAS because ABAS had several limitations; thus, we recently published a study comparing the performance of DLBAS with that of two commercially available ABAS systems for breast cancer RT.⁴⁵ In this study, the deep learning-based approach showed more consistent and robust performance than ABAS for most structures, and

this performance gap increased substantially for soft-tissue-based regions and smaller volumes.

DLBAS has been widely investigated in head & neck, lung, and prostate cancers, and has demonstrated clinically relevant impact regarding saving time and mitigating inter-observer variability.^{25,26,46} Although several studies have reported the feasibility of the deep learning-based approach for the breast, training and testing has only been performed for ipsilateral breast CTVs.^{47,48} In this study, a satisfactory DSC of 0.94 for CTVp_breast was shown, which is similar with that obtained for other series using the deep learning-based approach. One study using a dataset of 800 patients with a deep learning algorithm (DD-ResNet) showed a mean DSC of 0.91 for the CTVs of both breasts.⁴⁷ Furthermore, similar with the study by Eldesoky et al., in which they tested ABAS, we performed training not only for whole breast CTVs but also for various OARs and other CTVs, including regional lymph nodes in breast cancer patients. The current DLBAS model showed higher performance in segmenting various OARs—including heart, lung, thyroid, esophagus, and spinal cord—that were large and well-defined. As for regional lymph nodes, because of the smaller volumes and less well-defined borders, the current model exhibited modest performance. Regarding qualitative scoring, the expert and non-expert panels gave high difference and assistance scores for both the OARs and CTVs.

To date, even with fully validated auto-segmentations, modification or correction by clinical experts is commonly accepted. However, whether modification or correction is essential when auto-contours are utilized for dosimetric analysis has not been well studied. In this study, dosimetric analysis showed that there was a good agreement in dose distribution between manual

and auto-segmented contours for OARs. However, as for CTVs, particularly for the axillary lymph node regions and IMN, there were some discrepancies between manual and auto-segmented contours. For OARs, it can be suggested that auto-segmented contours can be used for dose-response related studies or predicting normal tissue complication probability in clinic. However, for CTVs, auto-contours in target volumes need significant modification by experts to conform to the corresponding anatomy and to individualize according to tumor and patient information. In the area of research, auto-segmented CTVs can be used as a reference point when comparing target volume delineation of various participants.

In breast cancer trials, variations in target delineation and RT planning have become a prominent issue, particularly in multidisciplinary trials that lack RT quality assurance programs.⁴⁹ In a recent audit study across a large network, it was found that nodes were not contoured or the contour quality was inadequate for 18% of patients.⁵⁰ In a Korean study that investigated inter-institutional variations in breast IMRT (KROG 19-01), there were large heterogeneities in the target volume as well as OARs, producing large variations in mean heart dose and lung V20Gy (up to five times in the same dummy run case). We believe that our auto-segmented contours of CTVs and OARs can play an important role in the breast RT quality assurance process, as illustrated in Chen et al.'s study.⁴⁸ Nationwide quality assurance is underway in Korea with our proposed algorithm.

Accurate delineation of all OARs and CTVs is a laborious task; here, auto-segmentation can serve as a useful tool in reducing the workload on physicians. In a previous study on ABAS for loco-regional RT of breast cancer, it was found that ABAS reduced the time required for manual segmentation

before correction by 93% and after correction by 32%.⁴⁴ This study showed a similar potential, with average times of 39 min and 10 min for manual delineation and DLBAS, respectively. With the assistance of DLBAS, radiation oncologists will be able to work more efficiently. Qualitative subjective scoring by the expert and non-expert panels exhibited satisfactory results for both difference and assistance scores, showing that DLBAS can serve as a helpful tool in real-world clinics.

C. EXTERNAL VALIDATION

Using our deep-learning based auto-contouring system, we compared the performance between manual contours from multiple experts, corrected-auto contours, and auto-contours in OARs for breast cancer radiotherapy. As a result, auto-contours were shown to have at least similar performance with manual contours of experts, shown by that average DSC and HD of auto-contours ranked second and first place, respectively, among the experts' manual contours. The inter-physician variation shown in manual contours was reduced with an aid of auto-contouring system. Moreover, the time for contouring was substantially reduced with an aid of auto-contouring system, with good user satisfaction.

Notably, in cases of breast contours, auto-contours or corrected auto-contours showed especially better accuracy than manual contours. The DSC values were improved by 0.09 (right breast) and 0.07 (left breast) in the corrected-auto-contours compared with the manual contours, while those with other OARs were mostly within 0.02. The HD values were also improved by 4.38 mm (right breast) and 2.71 mm (left breast), while those with other OARs were mostly within 2 mm. The hear contours also had better accuracy with the

auto-contouring system than manual contouring alone. In contrast, in cases of liver contours, HD values were prominently high in the auto-contours and reduced in the corrected-auto-contours to the similar value of manual contours, which means that manual adjustment was necessary. Therefore, the performance of manual contouring and auto-contouring are similar in general, but the detailed performance seems vary depending on the OARs.

We observed substantial inter-physician variability between the experts' manual contours. Substantial variability in manual contouring the targets and OARs between the institutions and observers has been shown in RTOG multi-institutional and multiple-observer study.¹⁴ Such inter-physician variability is obstacle in accurate assessment of the efficacy of radiotherapy and the risk of long-term side effect. Incidental radiation exposure to the heart during breast RT increases the risk of heart disease, considering the dose-response relationship between heart radiation dose and acute coronary event.^{10,51} Moreover, radiation-related hypothyroidism,⁹ radiation pneumonitis,⁵² and secondary contralateral breast cancer⁵³ have been reported in patients with breast cancer. In addition, in clinical trials including radiotherapy, there is a problem of standardization of treatment with this variability in delineating target and OARs.⁵⁴ In the RTOG 0617 trial, a radiation dose escalation trial in non-small cell lung cancer, an analysis using deep-learning segmented hearts showed that the actual heart doses were higher than originally reported due to the inconsistent and insufficient heart segmentation.⁵⁵ Our results showed that this problem can be solved by applying auto-contouring system. For example, the lateral border of breast had the largest variation among the experts' manual contours in this study. Because the most widely used RTOG⁵⁶ and ESTRO guidelines¹³ define the lateral border of breast as clinically palpable breast or lateral breast fold, it is difficult to clearly define it on CT image. The

auto-contouring system can help to standardize the delineation when the definition of the boundary of organ is ambiguous. Of note, a medical student who had no experience in breast cancer radiotherapy had bad performance in manual contouring but had comparable performance to other experts with an aid of auto-contouring system. This shows the potential that the learning can be sharper with an aid of auto-contouring system for training radiation oncology.

The manual adjustment of auto-contour had an average time reduction of 84% compared to the manual contouring. The time reduction was most remarkable in breast and liver contouring, which took the longest manual contouring time. When adjusting the auto-contour, the average time taken for each organ was less than 1 minute, which means that only minimal or no correction was needed. Adjusting liver and esophagus auto-contours took relatively longer time, but it was only 1.5 minutes, respectively. In addition, participants responded that auto-contouring system helped to shorten the time and would like to use it in the future. Considering that symptoms of burnout have been reported in >1 of every 2 practicing physician and this affliction,⁵⁷ which can be largely driven by work-related stressors,⁵⁸ efforts are needed to reduce the work loading of physicians, and ACS can be used for this.

Our study had several limitations. First, focusing on contours produced by a single expert had both positive and negative effects. One positive effect was that we were able to test each model's performance in a more controlled manner without introducing additional sources of variability. However, as mentioned earlier, we are aware that inter-observer variations exist and that this may affect the overall performance, so additional studies involving multiple experts will be necessary to test the robustness of our methods. Furthermore, since the findings of our study generated and validated by ground truth from a single expert,

further studies are required with external validation involving multiple experts. Finally, we did not optimize the ABAS for each software package. Since Mirada does not offer any options for algorithm adjustment per its policy, we simply sent the planning CT to the WFB server to generate the ABAS contours. Although MIM offers the tools for modifying and correcting the match between the template and the new subject during fusion and registration to ensure better ABAS outcomes, none of the edits were made in this study to maintain consistency with Mirada. Additionally, we carried out a simultaneous ABAS, where we segmented all the structures simultaneously. Due to a wide discrepancy in patient size, organ shape, and displacement, carrying out this procedure in a single step may not be the most accurate method.

V. CONCLUSION

Compared to ABAS, DLBAS was more consistent and robust in its performance across most structures. In this clinical study, we have confirmed the plausibility of these segmentation solutions for clinical implementations. This study also showed the potential and feasibility for DLBAS for breast cancer patients receiving RT after breast-conserving surgery. Although DLBAS cannot serve as a substitute for the experience of radiation oncologists, it has the potential to serve as a useful tool in assisting them. In addition, ACS showed at least similar performance in OARs compared with experts' manual contouring, which anticipates further applications of ACS to target volumes.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018;68:394-424.
2. Poortmans PM, Collette S, Kirkove C, Van Limbergen E, Budach V, Struikmans H, et al. Internal Mammary and Medial Supraclavicular Irradiation in Breast Cancer. *N Engl J Med* 2015;373:317-27.
3. Whelan TJ, Olivotto IA, Parulekar WR, Ackerman I, Chua BH, Nabid A, et al. Regional Nodal Irradiation in Early-Stage Breast Cancer. *N Engl J Med* 2015;373:307-16.
4. Thorsen LB, Offersen BV, Dano H, Berg M, Jensen I, Pedersen AN, et al. DBCG-IMN: A Population-Based Cohort Study on the Effect of Internal Mammary Node Irradiation in Early Node-Positive Breast Cancer. *J Clin Oncol* 2016;34:314-20.
5. Chang JS, Lee J, Chun M, Shin KH, Park W, Lee JH, et al. Mapping patterns of locoregional recurrence following contemporary treatment with radiation therapy for breast cancer: A multi-institutional validation study of the ESTRO consensus guideline on clinical target volume. *Radiother Oncol* 2018;126:139-47.
6. Chang JS, Byun HK, Kim JW, Kim KH, Lee J, Cho Y, et al. Three-dimensional analysis of patterns of locoregional recurrence after treatment in breast cancer patients: Validation of the ESTRO consensus guideline on target volume. *Radiother Oncol* 2017;122:24-9.
7. Byun HK, Chang JS, Im SH, Kirova YM, Arsene-Henry A, Choi SH, et al. Risk of Lymphedema Following Contemporary Treatment for Breast Cancer: An Analysis of 7617 Consecutive Patients From a Multidisciplinary Perspective. *Ann Surg* 2019.
8. Lee BM, Chang JS, Kim SY, Keum KC, Suh CO, Kim YB. Hypofractionated Radiotherapy Dose Scheme and Application of New Techniques Are Associated to a Lower Incidence of Radiation Pneumonitis in Breast Cancer Patients. *Front Oncol* 2020;10:124.
9. Choi SH, Chang JS, Son NH, Hong CS, Byun HK, Hong N, et al. Risk of Hypothyroidism in Women after Radiotherapy for Breast Cancer. *Int J Radiat Oncol Biol Phys* 2021.
10. Chung SY, Oh J, Chang JS, Shin J, Kim KH, Chun KH, et al. Risk of Cardiac Disease in Breast Cancer Patients: Impact of Patient-Specific Factors and Individual Heart Dose from Three-Dimensional Radiotherapy Planning. *Int J Radiat Oncol Biol Phys* 2021.
11. Gardner SJ, Kim J, Chetty IJ. Modern Radiation Therapy Planning and Delivery. *Hematology/Oncology Clinics* 2019;33:947-62.
12. Joosten A, Matzinger O, Jeanneret-Sozzi W, Bochud F, Moeckli R.

- Evaluation of organ-specific peripheral doses after 2-dimensional, 3-dimensional and hybrid intensity modulated radiation therapy for breast cancer based on Monte Carlo and convolution/superposition algorithms: implications for secondary cancer risk assessment. *Radiotherapy and Oncology* 2013;106:33-41.
13. Offersen BV, Boersma LJ, Kirkove C, Hol S, Aznar MC, Biete Sola A, et al. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer. *Radiother Oncol* 2015;114:3-10.
 14. Li XA, Tai A, Arthur DW, Buchholz TA, Macdonald S, Marks LB, et al. Variability of target and normal structure delineation for breast cancer radiotherapy: an RTOG Multi-Institutional and Multiobserver Study. *Int J Radiat Oncol Biol Phys* 2009;73:944-51.
 15. Chung Y, Kim JW, Shin KH, Kim SS, Ahn SJ, Park W, et al. Dummy run of quality assurance program in a phase 3 randomized trial investigating the role of internal mammary lymph node irradiation in breast cancer patients: Korean Radiation Oncology Group 08-06 study. *Int J Radiat Oncol Biol Phys* 2015;91:419-26.
 16. Ling DC, Moppins BL, Champ CE, Gorantla VC, Beriwal S. Quality of Regional Nodal Irradiation Plans in Breast Cancer Patients Across a Large Network-Can We Translate Results From Randomized Trials Into the Clinic? *Pract Radiat Oncol* 2021;11:e30-e5.
 17. Lee H, Lee E, Kim N, Kim JH, Park K, Lee H, et al. Clinical Evaluation of Commercial Atlas-Based Auto-Segmentation in the Head and Neck Region. *Front Oncol* 2019;9:239.
 18. Hwee J, Louie AV, Gaede S, Bauman G, D'Souza D, Sexton T, et al. Technology assessment of automated atlas based segmentation in prostate bed contouring. *Radiat Oncol* 2011;6:110.
 19. Pirozzi S, Horvat M, Piper J, Nelson A. SU-E-J-106: Atlas-Based Segmentation: Evaluation of a Multi-Atlas Approach for Lung Cancer. *Med Phys* 2012;39:3677.
 20. Lim JY, Leech M. Use of auto-segmentation in the delineation of target volumes and organs at risk in head and neck. *Acta Oncol* 2016;55:799-806.
 21. Boon IS, Au Yong TPT, Boon CS. Assessing the Role of Artificial Intelligence (AI) in Clinical Oncology: Utility of Machine Learning in Radiotherapy Target Volume Delineation. *Medicines (Basel)* 2018;5.
 22. Francolini G, Desideri I, Stocchi G, Salvestrini V, Ciccone LP, Garlatti P, et al. Artificial Intelligence in radiotherapy: state of the art and future directions. *Medical Oncology* (Northwood, London, England) 2020;37:50-.
 23. Oktay O, Nanavati J, Schwaighofer A, Carter D, Bristow M, Tanno R,

- et al. Evaluation of Deep Learning to Augment Image-Guided Radiotherapy for Head and Neck and Prostate Cancers. *JAMA Network Open* 2020;3:e2027426-e.
24. Fionda B, Boldrini L, D'Aviero A, Lancellotta V, Gambacorta MA, Kovács G, et al. Artificial intelligence (AI) and interventional radiotherapy (brachytherapy): state of art and future perspectives. *J Contemp Brachytherapy* 2020;12:497-500.
25. Kiljunen T, Akram S, Niemelä J, Löyttyniemi E, Seppälä J, Heikkilä J, et al. A Deep Learning-Based Automated CT Segmentation of Prostate Cancer Anatomy for Radiation Therapy Planning-A Retrospective Multicenter Study. *Diagnostics* 2020;10:959.
26. Brunenberg EJ, Steinseifer IK, van den Bosch S, Kaanders JH, Brouwer CL, Gooding MJ, et al. External validation of deep learning-based contouring of head and neck organs at risk. *Physics and Imaging in Radiation Oncology* 2020;15:8-15.
27. Jégou S, Drozdal M, Vazquez D, Romero A, Bengio Y. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2017. p.1175-83.
28. Offersen BV, Boersma LJ, Kirkove C, Hol S, Aznar MC, Sola AB, et al. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer. *Radiotherapy and oncology* 2015;114:3-10.
29. White J, Arthur D, Buchholz T, MacDonald S, Marks L, Pierce L, et al. Radiation Therapy oncology group breast cancer contouring Atlas. 2016.
30. Mir R, Kelly SM, Xiao Y, Moore A, Clark CH, Clementel E, et al. Organ at risk delineation for radiation therapy clinical trials: Global Harmonization Group consensus guidelines. *Radiotherapy and Oncology* 2020.
31. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention: Springer; 2015. p.234-41.
32. Tan M, Le QV. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 2019.
33. Milletari F, Navab N, Ahmadi S-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D Vision (3DV): IEEE; 2016. p.565-71.
34. Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, van der Stoep J, et al. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiother Oncol* 2018;126:312-7.

35. Ahn SH, Yeo AU, Kim KH, Kim C, Goh Y, Cho S, et al. Comparative clinical evaluation of atlas and deep-learning-based auto-segmentation of organ structures in liver cancer. *Radiat Oncol* 2019;14:213.
36. Jung JW, Lee C, Mosher EG, Mille MM, Yeom YS, Jones EC, et al. Automatic segmentation of cardiac structures for breast cancer radiotherapy. *Phys Imaging Radiat Oncol* 2019;12:44-8.
37. Dormer JD, Ma L, Halicek M, Reilly CM, Schreibmann E, Fei B. Heart Chamber Segmentation from CT Using Convolutional Neural Networks. *Proc SPIE Int Soc Opt Eng* 2018;10578.
38. Morris ED, Ghanem AI, Pantelic MV, Walker EM, Han X, Glide-Hurst CK. Cardiac Substructure Segmentation and Dosimetry Using a Novel Hybrid Magnetic Resonance and Computed Tomography Cardiac Atlas. *Int J Radiat Oncol Biol Phys* 2019;103:985-93.
39. Weber DC, Poortmans PM, Hurkmans CW, Aird E, Gulyban A, Fairchild A. Quality assurance for prospective EORTC radiation oncology trials: the challenges of advanced technology in a multicenter international setting. *Radiother Oncol* 2011;100:150-6.
40. Cox S, Cleves A, Clementel E, Miles E, Staffurth J, Gwynne S. Impact of deviations in target volume delineation - Time for a new RTQA approach? *Radiother Oncol* 2019;137:1-8.
41. Bridge P, Bridge R. Artificial Intelligence in Radiotherapy: A Philosophical Perspective. *J Med Imaging Radiat Sci* 2019;50:S27-s31.
42. Gentile MS, Usman AA, Neuschler EI, Sathiaselan V, Hayes JP, Small Jr W. Contouring guidelines for the axillary lymph nodes for the delivery of radiation therapy in breast cancer: evaluation of the RTOG breast cancer atlas. *International Journal of Radiation Oncology* Biology* Physics* 2015;93:257-65.
43. Wong J, Fong A, McVicar N, Smith S, Giambattista J, Wells D, et al. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. *Radiotherapy and Oncology* 2020;144:152-8.
44. Eldesoky AR, Yates ES, Nyeng TB, Thomsen MS, Nielsen HM, Poortmans P, et al. Internal and external validation of an ESTRO delineation guideline - dependent automated segmentation tool for loco-regional radiation therapy of early breast cancer. *Radiother Oncol* 2016;121:424-30.
45. Choi MS, Choi BS, Chung SY, Kim N, Chun J, Kim YB, et al. Clinical evaluation of atlas-and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. *Radiotherapy and Oncology* 2020.
46. Poortmans PM, Takanen S, Marta GN, Meattini I, Kaidar-Person O. Winter is over: The use of Artificial Intelligence to individualise

- radiation therapy for breast cancer. *The Breast* 2020;49:194-200.
47. Men K, Zhang T, Chen X, Chen B, Tang Y, Wang S, et al. Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning. *Phys Med* 2018;50:13-9.
48. Chen X, Men K, Chen B, Tang Y, Zhang T, Wang S, et al. CNN-Based Quality Assurance for Automatic Segmentation of Breast Cancer in Radiotherapy. *Front Oncol* 2020;10:524.
49. Peters LJ, O'Sullivan B, Giralt J, Fitzgerald TJ, Trotti A, Bernier J, et al. Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02. *Journal of clinical oncology* 2010;28:2996-3001.
50. Ling DC, Moppins BL, Champ CE, Gorantla VC, Beriwal S. Quality of Regional Nodal Irradiation Plans in Breast Cancer Patients Across a Large Network-Can We Translate Results From Randomized Trials Into the Clinic? *Pract Radiat Oncol* 2020.
51. Darby SC, Ewertz M, McGale P, Bennet AM, Blom-Goldman U, Brønnum D, et al. Risk of ischemic heart disease in women after radiotherapy for breast cancer. *N Engl J Med* 2013;368:987-98.
52. Choi J, Kim YB, Shin KH, Ahn SJ, Lee HS, Park W, et al. Radiation Pneumonitis in Association with Internal Mammary Node Irradiation in Breast Cancer Patients: An Ancillary Result from the KROG 08-06 Study. *J Breast Cancer* 2016;19:275-82.
53. Zhang Q, Liu J, Ao N, Yu H, Peng Y, Ou L, et al. Secondary cancer risk after radiation therapy for breast cancer with different radiotherapy techniques. *Sci Rep* 2020;10:1220.
54. Perez CA, Gardner P, Glasgow GP. Radiotherapy quality assurance in clinical trials. *Int J Radiat Oncol Biol Phys* 1984;10 Suppl 1:119-25.
55. Thor M, Apte A, Haq R, Iyer A, LoCastro E, Deasy JO. Using Auto-Segmentation to Reduce Contouring and Dose Inconsistency in Clinical Trials: The Simulated Impact on RTOG 0617. *Int J Radiat Oncol Biol Phys* 2020.
56. Breast cancer atlas for radiation therapy planning: consensus definitions. Available. Available at <https://www.nrgoncology.org/ciro-breast> [Accessed Jan 1, 2021]
57. Shanafelt TD, Hasan O, Dyrbye LN, Sinsky C, Satele D, Sloan J, et al. Changes in Burnout and Satisfaction With Work-Life Balance in Physicians and the General US Working Population Between 2011 and 2014. *Mayo Clin Proc* 2015;90:1600-13.
58. Dyrbye LN, Burke SE, Hardeman RR, Herrin J, Wittlin NM, Yeazel M, et al. Association of Clinical Specialty With Symptoms of Burnout and Career Choice Regret Among US Resident Physicians. *Jama*

2018;320:1114-30.

ABSTRACT (IN KOREAN)

유방암 환자 방사선 치료의 정상 장기 및 치료 체적의 자동 구획화

<지도교수 금 기 창>

연세대학교 대학원 의학과

장 지 석

목적: 유방암 방사선 치료에서 치료 체적에 대한 정확한 타겟 그리기는 중요하다. 하지만 방사선 치료 계획 과정에 타겟 그리기는 의료진의 부담을 주고 있으며, 의료진 간의 편차는 존재하고 있다. 본 연구에서는 Deep learning-based auto-segmentation (DLBAS)의 성능을 atlas-based segmentation solutions (ABAS)와 비교하고, 임상 의사의 관점에서 유용성을 평가하고, 최종적으로 외부 타당도 조사를 통하여 유방암 방사선 치료에서 자동 구획화의 가능성을 규명하고자 한다.

대상 및 방법: 유방암 방사선 치료 체적과 정상장기들에 대하여 한 명의 연구진에 의하여 구획화 정보를 생성하였다. Convolutional neural network 알고리즘을 이용하여 auto-contours를 생성하였고, Dice similarity coefficient (DSC) and 95% Hausdorff distance (HD)를 이용하여 ABAS와 비교하였다. DLBAS에 의해 생성된 auto-contours의 질적인 평가를 조사하였고, manual contours와 방사선 치료 선량-체적 히스토그램을 비교하여 주요 선량평가분석을 시행하였다. 마지막으로 2개 기관의 11명의 전문가에게 manual contour를 그릴 것을 요청하여 데이터를 수집하였다. 외부 위원회를 통해 가장 최적의 치료 체적을

선정하였고, 나머지 10명의 contour와 DLBAS에 의해 생성된 auto-contour의 성능을 비교하여 순위 평가를 시행하였다.

결과: 제안된 DLBAS 모델은 대부분의 체적 (특히, 치료 체적과 심장 세부구조)에서 ABAS보다 더 일관된 결과와 높은 DSC와 낮은 HD 결과 값을 보였다. ABAS는 연조직의 정상장기와 조영제를 쓰지 않은 새로운 데이터 셋에서 DLBAS에 비해, 제한적인 성능을 보였다. 질적 평가를 위한 설문조사가 시행되었고, 중위수 8점으로 manual contour와 auto-contour 사이의 차이가 크지 않다고 대답하였으며, 임상에서 도움이 될 것으로 답변하였다. 또한 선량평가 분석 결과에서 차이는 미미하였다. 외부 검증 결과, 9개의 정상장기를 그리는데 평균 37분이 걸렸고, DLBAS는 6분이 걸렸다. Auto-contour는 전체 12개 중 1위 manual contour와 비교하였을 때 가장 DSC상 차이가 적었으며, HSD상 2번째로 차이가 적었다. 정상장기에서 가장 편차가 높았던 부위는 유방이었다.

결론: 유방 방사선 치료 계획에서 DLBAS의 실현가능성은 이번 연구에서 다각도로 검증되었다. 의료진의 최종 수정 과정은 필수적이지만, 앞으로 DLBAS는 방사선 치료를 도울 수 있는 훌륭한 가능성을 보여주었다.

핵심되는 말: 유방암, 자동 구획화, 방사선 치료, 인공지능

PUBLICATION LIST

Chung SY, Chang JS, Choi MS, Chang Y, Choi BS, Chun J, et al. Clinical feasibility of deep learning-based auto-segmentation of target volumes and organs-at-risk in breast cancer patients after breast-conserving surgery. *Radiation Oncology* 2021;16:44

Choi MS, Choi BS, Chung SY, Kim N, Chun J, Kim YB, et al. Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. *Radiotherapy and Oncology* 2020;153:139-45.